BARLOW, H. B., & FÖLDIÁK, P. (1989). Adaptation and decorrelation in the cortex.
In R. Durbin, C. Miall, & G. Mitchison (Ed.),
The Computing Neuron (pp. 54-72). Wokingham, England: Addison-Wesley.

4
Adaptation and Decorrelation in the Cortex

Horace Barlow & Peter Földiák

Summary

Any small region of the cortex receives input through a large number of afferent fibres, and transmits efferent output to other regions of the brain. If the units interact according to an anti-Hebbian rule, the outputs define a coordinate system in which there are no correlations even when the input fibres show strong correlations. The idea that cortex performs such decorrelation has several theoretical merits and fits some prominent facts about the cortex:

1) It would be advantageous in making effective use of the narrow dynamic range that is characteristic of cortical neurons.
2) The absence of correlations would make it easier to detect newly appearing associations resulting from new causal factors in the environment.
3) It provides a role for recurrent collaterals, which are a conspicuous feature of cortical neurons, especially in striate areas.
4) It may account for the after-effects of adapting to patterned stimuli.
5) It could account for part, at least, of the effects of experience during cortical development.
6) It can improve the performance of Hebbian learning rules in associative networks.

4.1 How is the cortical image interpreted?

Many readers will be familiar with the picture of neural activity in the primary visual cortex that has been revealed by the work of Hubel (1988), Wiesel and many others over the past two or three decades, but most will probably agree that nobody knows quite what will happen next.

Let us put the problem this way: what can these signals for lines, edges, textures, movements, disparities and colours mean without any background

knowledge of the world? Are they not like single letters without a language? And without knowing what effects they produce, the occurrence of activity in a nerve cell does not mean any more than the production of silver grains means to the camera. You have to know something about the interpretive system before you can understand what the nerve impulses mean.

In this paper we are concerned with one particular aspect of this system, namely the way it acquires, stores, and uses background knowledge of the sensory environment. Probably everyone will agree that visual perception uses knowledge of images and their ways to reach a valid interpretation of the current scene; to illustrate this, ask yourself when you last tripped, or avoided tripping, over a shadow. According to current knowledge a shadow will stimulate an 'edge-detector' in V1 as effectively as the image of a doorstep, so why is it that you do not treat shadows as if they were doorsteps all the time? Possibly current knowledge is incomplete and it is certainly worth asking how our interpretive system distinguishes them.

There are two groups of people who are acutely aware of the major task facing the interpretive system in going from the 2-D images our eyes provide, to the world of 3-D objects that we can usually construct with astonishing reliability. The first are the perceptual psychologists, and few have appreciated the problem from this point of view better, or more deeply, than did Helmholtz nearly a century and a half ago (Helmholtz, 1925); that fact perhaps indicates how slow and difficult progress has been using classical psychological methods. Much more recently, those concerned with robot vision and computer image interpretation have tried to mimic our everyday perceptual performance - and having so far failed, they are very impressed by what our brains do so effortlessly.

One of the basic problems, understood since Helmholtz's time and formalized by Poggio, Torre & Koch (1985), is that 2-D images do not provide all the information that is needed for the reconstructions that our minds perform. Assumptions are necessary to make this possible, and if these assumptions are to work they must be based on valid expectations about the nature of the world we inhabit. How are these expectations formed? What are they based upon? How are they stored? How are they accessed when needed? We think that well-known adaptation phenomena provide strong clues to the solution of this puzzle.

4.2 Adaptation to patterns

Aristotle described the waterfall phenomenon, now known as the motion after-effect: you gaze at a waterfall for about a minute, then when you transfer your gaze to the rocks or foliage at the side they appear to be moving upwards. The idea that this may be something to do with entraining eye movements can be ruled out by using a rotating spiral. The effect is easily visible, though the after-movement is not very fast or vigorous - a sort of drift. It is confined to the region of the visual field exposed to the adapting movement.

The explanation usually given, and actually dating back to Sigmund Exner (1894), is that the period of adaptation fatigues elements responding selectively to motion in a particular direction, and it has been shown that something of this sort does happen with directional and other types of pattern selective cells (Barlow & Hill,1963; Maffei, Fiorentini & Bisti, 1973; Vautin & Berkeley, 1977; Ohzawa, Sclar & Freeman, 1985). When you continue stimulating such a cell the response declines; then when you stop, the maintained discharge is suppressed and its responsiveness is impaired. There is usually no adapting effect from stimuli to which the cell does not respond. Thus the explanation offered, for instance for the motion after-effect, is that continued motion in one direction imbalances the responses from units signalling opposite directions, so that a stationary, non-moving, object appears to move in the reverse direction.

Similar adaptation effects can be obtained with many patterned stimuli. Gibson (1933, 1937) noted after-effects from viewing curved lines, and Gilinsky (1968), Pantle & Sekuler (1968) and Blakemore & Campbell (1969) independently discovered that a plain grating causes decreased sensitivity to gratings of similar frequency and orientation. As shown in Figure 4.1, adapting to coarsely spaced lines makes lines appear more finely spaced, and vice versa (Blakemore & Sutton, 1969), and tilted lines cause lines to appear tilted the other way.

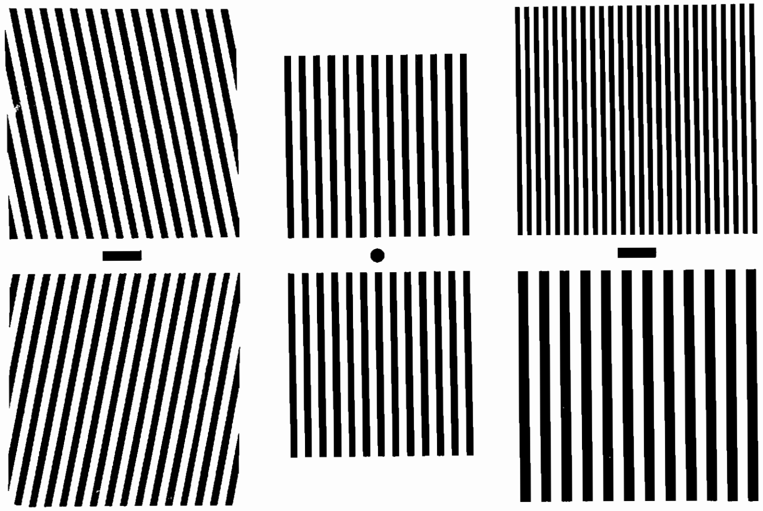

Figure 4.1 The visual system changes its characteristics when it adapts to grating patterns. Look at the black bar between the pair of tilted gratings for about 30 seconds, then shift your gaze to the round dot between the central pair of vertical gratings: they will appear tilted in the opposite direction to that to which the corresponding region of the visual field has just been adapted. Similarly the apparent spacing of the gratings can be changed by adapting to the pair of gratings on the right (from Blakemore, 1973).

Finally there is the mysterious McCollough (1965) effect. Adaptation consists of looking at a red, horizontal, grating for about 5 s, then shifting your gaze to a green, vertical, grating for another 5 s, then back to the red and so on for a few minutes. When you then look at a test pattern consisting partly of horizontal, partly of vertical, colourless gratings, you will see the horizontal parts tinted green, the vertical parts red. If you think the test pattern itself is tinted, tilt your head through 90°, and observe that the tints reverse.

The important point about the McCollough effect is that it persuaded people to realize that one is adapting to *contingencies* (red if horizontal, green if vertical), rather than to a primary sensory quality (for a review, see Stromeyer 1978). The issue is blurred by the fact that there probably are neurons selectively sensitive to these contingencies (Michael, 1978), so that the 'fatigue of pattern selective neurons' explanation still holds water. But there are at least three reasons to think there is more to adaptation than this.

We first warn against the assumption that, because a stimulus is effective in activating a single neuron, it is that activation alone that causes change in the neuron's responsiveness. The adapting stimuli will generally excite large numbers of neurons having the same type of pattern specificity, and we think it is this joint excitation that is the important contingency for causing adaptation. Thus for the loss of sensitivity that follows viewing a plain grating the contingency would be the joint excitation of many neurons all tuned to the same orientation and spatial frequency; likewise for the disturbances in perception illustrated in Figure 4.1. Though the grating would certainly be an effective stimulus for many individual neurons, we think it is the joint excitation of a group of them, and the resulting changes in their interactions, that causes changes in their responsiveness. A mechanism that might bring about such group effects will be proposed shortly.

The second point is that images are full of non-accidental associations; although we make the phenomenon observable by introducing *abnormal* contingencies, the *normal* patterns of contingent excitation must continually produce their effects on the mechanism. If adaptation to these naturally occurring contingencies is a mechanism producing positive functional effects, it should not be called fatigue.

The third point follows from this: the naturally occurring contingencies are precisely what the interpretive mechanism needs to know about to provide valid expectations upon which reliable interpretations of 2-D images depend. This is so because these natural contingencies result from the properties of real objects in the real three dimensional world, and this is what provides the eye with its normal diet of images. We therefore think the adaptation mechanism is fundamentally important for acquiring, storing and accessing valid knowledge about the normal environment, and hence 'fatigue' is a thoroughly inappropriate term for it.

One can look at this from many different points of view, and we shall first present it as a necessary mechanism, given the poor dynamic range of cortical neurons. Next we shall suggest a physiological process that might bring it about and an idealized model of this process. Finally, we shall return to a discussion at the perceptual level.

4.3 The need for contingent adaptation

Most people are familiar with the idea that normal light adaptation adjusts the response range of neurons to the range of luminances actually present in the image. Figure 4.2 shows at the top a hypothetical distribution of luminances over an image at a low light level, and another at high. Below are shown hypothetical response curves of 'on' and 'off' signalling neurons, and how they shift with mean luminance of the image.

Something of this sort really does happen; Figure 4.3 shows responses of two retinal ganglion cells in the cat. The intersections of the pairs of curves with the abscissa show the levels to which the retina was adapted; the ordinates show how many extra impulses were elicited when the luminance at the centre of the receptive field was briefly stepped up or down to other neighbouring values on the abscissa. Clearly the response range shifts with adapting luminance, and this is necessary because the range of possible input luminances is very large, while the range of possible outputs is strictly limited.

Now the available response range of neurons can be wasted in a different way. Suppose that, for whatever reason, two neurons very nearly always respond together; the available response space spanned by the two neurons will not be properly utilized, and Figure 4.4 shows what can be done about it. The top left shows a hypothetical plot of the response Ψ_A of one neuron to physical variable A against the response Ψ_B of another neuron to physical variable B. For simplicity, we have assumed uniform, rectangular distributions for each of them. If they are uncorrelated, the scatter diagram fills the whole plane uniformly. If each neuron can discriminate 4 levels of activity, all 16 possible states occur with approximately equal probabilities, and the limited representational possibilities of these two neurons would be fully utilized. The results in Barlow *et al.* (1987) suggest that four reliably distinguishable levels may be an optimistic estimate of the dynamic range of cortical neurons.

Next suppose the physical variables A and B are positively correlated, so the scatter diagram would then look like the one at top right. Clearly it would be inefficient for the two neurons to respond simply to physical variables A and B as before, because their responses would then be strongly correlated and the representational space would not be filled. An oblique coordinate system should be adopted so that the response Ψ_A is given by the projection on Ψ_A parallel to the Ψ_B axis, and vice versa. The diamond-shaped response pattern would now match the scatter of the points, and all representational possibilities would again be utilized.

At bottom right the distribution is plotted with the responses Ψ_A, Ψ_B as the orthogonal axes, and the dotted vectors show the directions of the original physical stimuli A and B in this plot. The vector for A slopes backwards, and this means physical stimulus A has a negative or inhibitory effect on Ψ_B, the neuron that originally responded only to B. Similarly B has an inhibitory effect on Ψ_A.

To summarize the argument of this section, when two physical variables are strongly correlated one might expect two neurons, each of which responds predominantly to one of them, to develop mutual *repulsion* so that if a physical

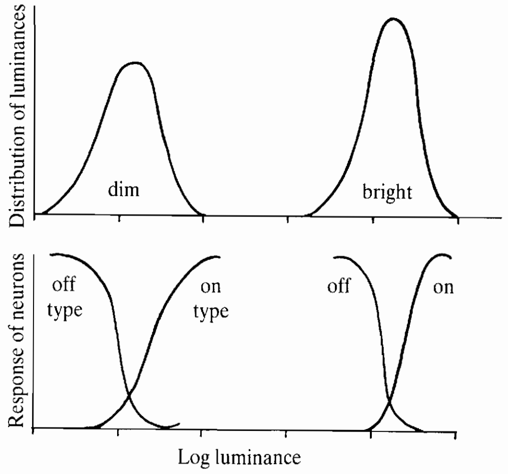

Figure 4.2 This illustrates the way retinal ganglion cell response characteristics change when the mean luminance of a scene changes. The top pair shows hypothetical distributions of luminance in a dim and a bright scene. Below are shown the hypothetical responses of retinal neurons to the luminances in the scenes; on-type neurons respond to increases and off-type to decreases, but without a shift in their characteristics when the mean luminance changes, the signals would not cover the appropriate luminance ranges.

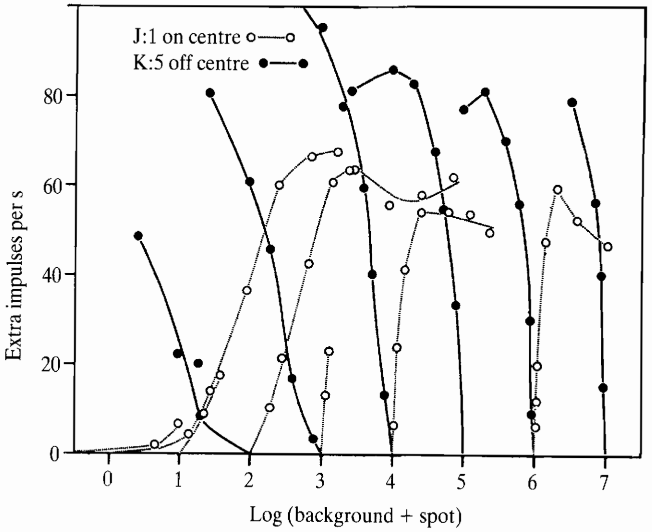

Figure 4.3 Actual retinal neurons behave as illustrated diagrammatically in Figure 4.2. The retina was adapted to the 7 different luminances at the points on the abscissa intersected by the curves. The luminance at the centre of the receptive field was then transiently shifted upwards for the on-centre unit (solid lines), or downwards for the off-centre unit (dashed lines), and their responses for such shifts of luminance are plotted as ordinates (from Barlow, 1969).

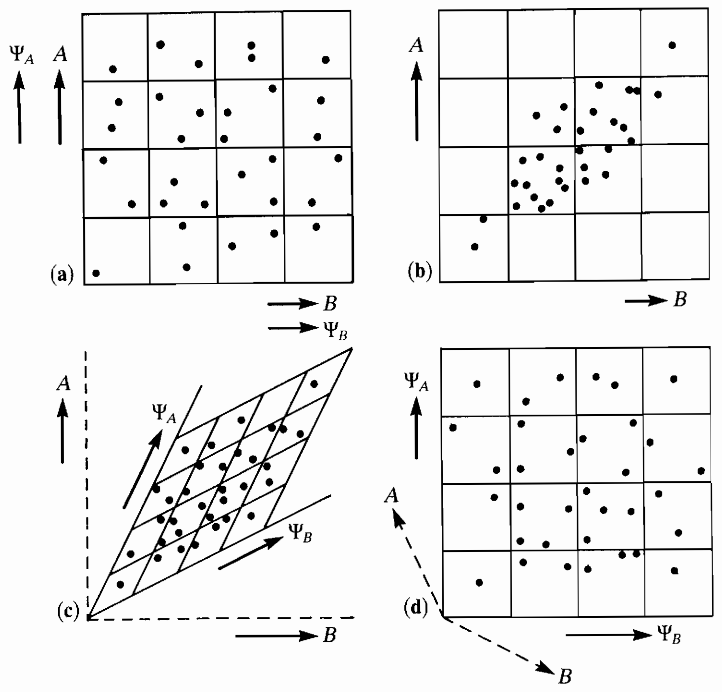

Figure 4.4 The limited dynamic range of neurons would be better used if their responses were uncorrelated. Such a state is shown in (a), where two uncorrelated physical stimuli A and B cause responses Ψ_A and Ψ_B in two neurons: all the distinguishable responses occur, as indicated by the presence of dots in all the squares. In (**b**) the physical stimuli are strongly correlated, and many of the distinguishable response states do not occur, indicated by empty squares. The solution is to make the response axes oblique, as in (**c**), where the response Ψ_A is given by the projection of a point on the oblique axis Ψ_A in the direction given by Ψ_B ; the lozenge shaped region now fits the pattern of correlated signals and all possible signals are generated. In (**d**) the orthogonal axes are the responses of the two neurons, and the dashed lines show the directions of the physical stimuli A and B; it will be seen that A slopes backwards, signifying that physical stimulus A has an inhibitory effect on the response Ψ_B, and similarly B inhibits Ψ_A.

variable excites the one it will inhibit the other, and vice versa. Such mutual inhibition is of course a well known feature of sensory ganglia, and the only new suggestion here is that its strength should be variable and adjusted according to the strength of the correlation between the activities of the two neurons under consideration. This argument stems simply from the desirability of utilizing fully the representational space offered by elements of limited dynamic range.

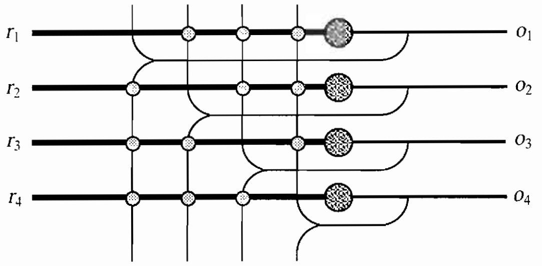

Figure 4.5 The idealized linear network with outputs o_i feeding back to the inputs r_i through modifiable connections; the strength of mutual inhibition increases when outputs are positively correlated. In matrix notation $\mathbf{o} = \mathbf{r} + \mathbf{Wo}$, so after an initial transient, $\mathbf{o} = (\mathbf{I} - \mathbf{W})^{-1}\, \mathbf{r} = \mathbf{T}\, \mathbf{r}$, where $\mathbf{T} = (\mathbf{I} - \mathbf{W})^{-1}$ is the overall transfer matrix of the network.

4.4 An idealized model

We are suggesting that there can be two kinds of adaptation: each single unit adapts to the mean values of its own input, while an inter-unit mechanism stores and discounts expected relationships between input variables. To demonstrate the effect of this mechanism we considered a simple model similar to Kohonen's novelty filter (1984). It consists of a number of linear units with inputs r_i, outputs o_i, and modifiable feedback connections between them (Figure 4.5).

The output of each of the N units is determined by the external input to that unit plus the feedback it receives from the other units, i.e. the outputs of the other units weighted by the synaptic strengths.

$$o_i = r_i + \sum_{j=1}^{N} w_{ij} o_j \qquad (4.1)$$

Initially the synapses are not effective, $w_{ij} = 0$. The network is trained by repeatedly taking patterns out of a predetermined, large set and presenting them on the input lines. The outputs will satisfy the above equation after an initial transient has occurred, which is not simulated.

For the definition of the learning rule it seemed convenient (but may not be necessary) to use a scaled version of the output variables, O_i, defined to have unit variance ($\langle O_i^2 \rangle = 1$), and calculated as $O_i = o_i \big/ \sqrt{\langle o_j^2 \rangle}$, where $\langle\ \rangle$ denotes averaging over the set of patterns. After reaching a stable output for each input pattern, the strengths of the synapses change very slightly according to the following symmetric, anti-Hebbian rule:

$$\begin{array}{ll} \Delta w_{ij} = -\,\alpha\, O_i\, O_j & \text{if } i \neq j; \\ \Delta w_{ii} = 0 & \text{otherwise} \end{array} \qquad (4.2)$$

where α is a small positive constant determining the rate of adaptation. If α is small enough, the synaptic weights between two units will change in (negative) proportion to the correlation between the output activities of the units taken over the whole input set. This means that if two output variables are initially positively correlated, then inhibition between them will gradually get stronger, making them less correlated, and the network can only reach a stable state ($\langle \Delta w_{ij} \rangle = 0$) when the outputs of all pairs of units are uncorrelated: $\langle O_iO_j \rangle = 0$ if $i \neq j$, and as $\langle O_i^2 \rangle = 1$, the correlation matrix tends to the identity matrix. The modification rule is unsupervised and local, the information about the activity of the pre- and post-synaptic units being available at synapse w_{ij}.

Figure 4.6 shows a correlated input distribution (each dot representing an input vector) and the output of a network consisting of two units after a few cycles of training.

The speed of convergence decreases with the number of units, but as Figure 4.7 shows, the covariance matrix of the outputs approaches the identity matrix at a nearly exponential rate.

4.5 Relation to Kohonen's novelty filter

The network and modification rules are the same as the novelty filter described by Kohonen & Oja (1976), with one minor difference; our output variables do not feed back on themselves, but are gain-controlled to have unit variance. As a result the output does not go to zero for a 'familiar' input. We also look upon the filter as performing a somewhat different task. Instead of learning a specific set of inputs (the number of which is smaller than the number of units) and signalling the projection of the current input on the complementary subspace of the set of learned vectors, we regard our network as learning the average relations between a large ensemble of inputs - its covariance matrix - and using this knowledge to generate uncorrelated signals. The number of pattern vectors can be arbitrarily large (>> N), or the input can be thought of as a continuous function of time.

4.6 Decorrelating taste signals

It is known that the information about the four basic tastes (sweet, sour, bitter, salty) is not carried by separate fibres; instead each fibre carries a mixed signal with different relative sensitivities to the four substances (Pfaffman, 1941; Sato, 1980). This lack of segregation has always been puzzling, partly because it contrasts with other sensory pathways, but also because the task of separating the mixed-up taste qualities does not appear a simple one. The puzzle would be lessened if neural networks can easily decorrelate sensory signals along the lines we have described, so we have tried out our algorithm on a simulated taste system.

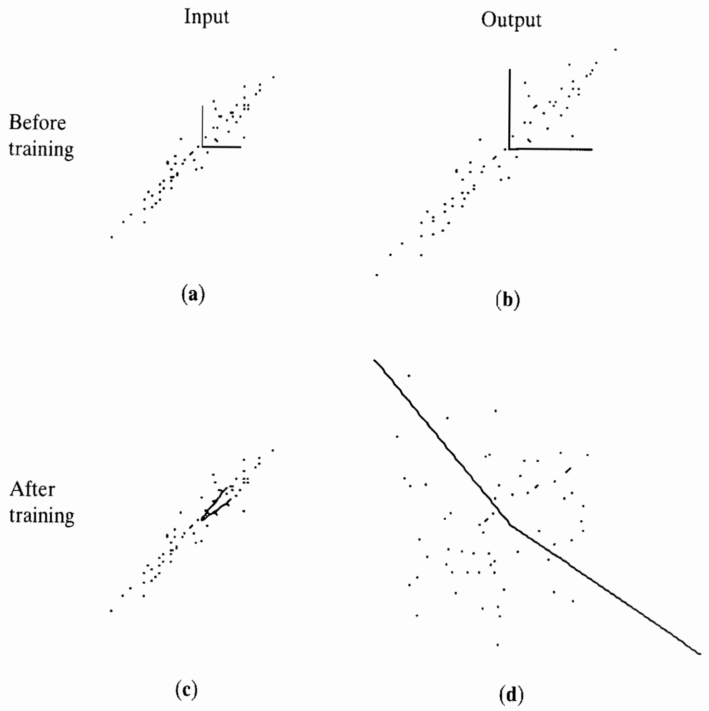

Figure 4.6 The dots represent a correlated set of vectors on (a) the inputs and (b) the outputs of a network consisting of two units before training. After training these are transformed to (c) and (d). Axes in the output space indicate the projections of the base vectors of the input space and vice versa.

If we assume independent distributions for the concentrations of the four taste substances to which the animal is exposed, the inputs to the units will be correlated because of the overlap in the relative sensitivities between any two units. During simulated training the outputs of the network become uncorrelated, but this does not mean that we get our original variables back. For instance, if the external signals S (sour) and B (bitter) are originally uncorrelated with normal distributions of equal variance, then S+B and S–B are also uncorrelated, as is any other rotation of the coordinates. Depending on the way the signals are mixed on the inputs the network will implement the transformation that can be achieved by a symmetric feedback matrix. If one wants to get not only an uncorrelated combination of variables but the original variables themselves, one has to assume a special property of the original signals. For instance, the concentrations of the four basic taste substances can only take positive values and in this case the only transformations that will keep all the output values positive (and uncorrelated) will be the ones that restore the original variables. Any other transformation will cause negative components in the outputs for some vectors.

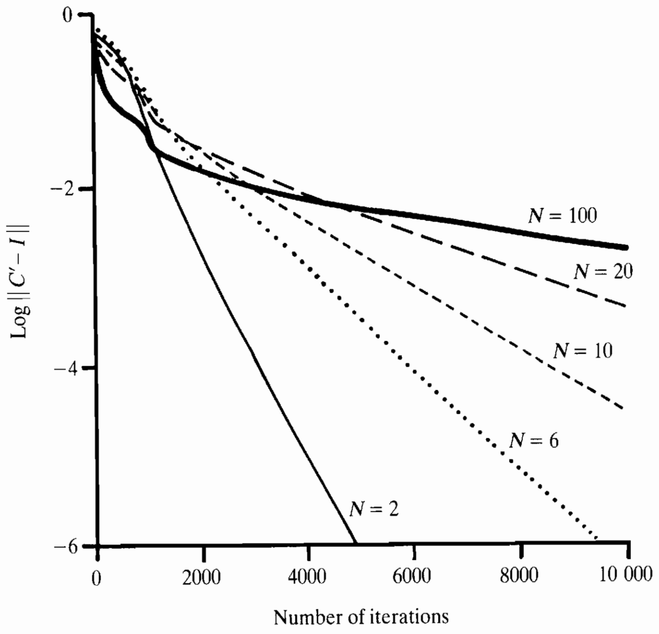

Figure 4.7 The rate of adaptation of the decorrelating network. For the purpose of estimating the speed of the convergence of the network, the input is assumed to have a Gaussian distribution, so it is not necessary to represent the input distribution as a set of patterns and its covariance matrix **V** uniquely defines it. For each input distribution the time of convergence will be different, so it can be estimated by averaging over a large number of randomly chosen input distributions. The covariance matrix of the outputs, **C**, can be calculated from each input covariance matrix, **V**, as: $\mathbf{C} = \mathbf{T}\,\mathbf{V}\,\mathbf{T}^{\mathrm{T}}$, which converges to the unit matrix during training. Single-unit adaptation is modelled by its effect on the covariance of the scaled output, $C'_{ii} = 1$, so $\mathbf{C}' = c_{ij} / \sqrt{c_{ii}\, c_{jj}}$. Initially the connections are ineffective, $w_{ij}{}^{(0)} = 0$, and they change as:

$$w_{ij}{}^{(k+1)} = w_{ij}{}^{(k)} - \alpha\, C'_{ij}{}^{(k)} \text{ if } i \neq j$$

after k cycles of training. Let's denote the expression $\frac{1}{N}\sqrt{\sum_{ij}\left(a_{ij} - b_{ij}\right)^2}$ by $\|\,\mathbf{A} - \mathbf{B}\,\|$ for matrices **A** and **B**. To measure the precision of the convergence we can use the quantity $\|\,\mathbf{C}' - \mathbf{I}\,\|$ to indicate how close the covariance matrix of the output is to the unit matrix. The diagram shows these values averaged over 100 runs for $N = 2, 6, 10, 20$ and averaged over two runs for $N = 100$. The random input covariance matrices were generated as: $\mathbf{V} = \mathbf{M}\mathbf{M}^{\mathrm{T}}$ where the elements of **M** were taken from an even distribution on the interval [0,1]. α was 0.001 for this simulation, and for this value 1 out of 100 runs oscillated and only converged with a smaller α. This one was replaced by an additional run. For higher values of α the convergence gets significantly faster, but a higher proportion of runs will oscillate.

A modified version of the model uses this assumption that the variables cannot take negative values in order to restore the original variables. Here the concentrations are taken from exponential distributions, chosen because these are simple distributions limited to positive values, and the synaptic modification rule is slightly modified:

$$\Delta w_{ij} = -\,\alpha\,(\,g(O_i)\,O_j - 1\,) \quad \text{if } \mathrm{i} \neq \mathrm{j} \tag{4.3}$$

where

$$g(x) = \left\{ \begin{array}{ll} x & \text{if } x \geq 0 \\ d{\cdot}x & \text{if } x < 0 \end{array} \right\} \tag{4.4}$$

and $d > 1$. Note that here O_i denotes the output scaled so that $\langle\, O_i \,\rangle = 1$ (rather than $\langle\, O_i^2 \,\rangle = 1$).

The zero point of our model variables (O_i) is assumed to correspond to the spontaneous firing rate of real neurones, and small negative values to firing below spontaneous rate. This rule behaves like the previous one as long as all outputs stay positive, but as soon as one of the O_i's goes slightly negative (meaning 'no firing') for one of the input patterns the negative value of that unit gets multiplied by d, a large positive constant, making Δw_{ij} positive. This results in a decrease of inhibition on the unit that went negative until all the patterns occupy the positive region of the output space only, restoring the original variables sweet, sour, bitter and salty.

It is known that in higher stages of the taste system some individual cells become more selective to only one of the basic tastes (Yamamoto *et al.*, 1985; Scott *et al.*, 1986), and our model shows how this might be achieved. When decorrelating many different groups of fibres in parallel, it would be advantageous to have the same coordinate system for each group, and having units selective for the primary qualities might, for instance, enable a salt-hungry animal to recognize what it lacks. Of course there is more to taste than correcting the confusion caused by the initial failure of the system to segregate the qualities on to separate pathways.

4.7 Local decorrelation in images

A decorrelating network can easily handle more than the four inputs shown in the model, but it cannot handle anything like the number that are involved in the images coming from the eyes. Furthermore the local connections in the striate cortex, which would be involved in the mutually inhibitory or excitatory connections between neurons, only extend for distances of the order of millimetres. Hence one must consider what decorrelation performed on local patches of the image could achieve, and first we need to clarify the nature of the information stored.

Note that simple automatic gain controls on each of the input lines might be said to store the mean values of each input, for the settings of the gains tell one how to derive a standard output level from each input. What is the information about the input that is stored in the model's matrix of inhibitory coefficients?

Without inhibition the outputs would be correlated in the same way as the inputs, while with the appropriate inhibitory coefficients the outputs are uncorrelated. Since the matrix of inhibitory coefficients compensates for the input correlations, it can be said that they store knowledge of the covariance matrix of the input in the normal environment. The pattern of saccadic eye-movements over a scene will cause the covariances to be averaged for different positions, but the values for different separations and orientations will be preserved and effectively yield the auto-correlation function. From this one can obtain the complete local power spectrum of the input images by Fourier transformation. Hence a decorrelating network stores knowledge of the power spectrum of its input and will be expected to show adaptive changes when exposed to images that have mean local power spectra or auto-correlation functions that differ from those to which they have previously been adapted.

Since decorrelation depends upon the local power spectrum, it is easy to see that many of the adaptation effects we have described are exactly what would be expected. What is interesting, however, is that this explanation does *not* at first sight seem to entail elements that are specifically tuned to different spatial frequencies, as in the usual 'channels' explanation (Braddick *et al.*, 1978). A decorrelating network would, it seems, necessarily adapt to stimuli that cause peaks in the local power spectrum, and would proceed to flatten them, simply because, with the moving eye, it is the power spectrum that drives the decorrelation process. On this view adaptation does not depend on the selectivity of the output neurons in the way that is often assumed, so one should be very cautious in accepting the existence of selective adaptation as evidence for neurons with specific selectivity for the adapting stimulus.

Notice that a network decorrelating N inputs uses of the order of N^2 synapses to store the power spectrum, which only requires N coefficients for its specification. As Kohonen (1984) points out, a sparse pattern of feedback connections should be sufficient to do this, or alternatively one might devise a means of storing more about the input in a completely connected network: for instance the different synapses might use different updating coefficients to produce different adaptation rates, or the feedback paths might be delayed by varying times in order to decorrelate temporal as well as spatial associations.

4.8 Temporal aspects

The normal environment is presumed to be whatever has happened in the recent past, and something needs to be said about how fast these normal expectations are formed. The evidence about this from experiments on pattern adaptation is curious. Fifteen seconds adaptation produces quite easily detectable after-effects, which usually persist for only a short time. Very thorough adaptation, on the other hand, can produce after-effects that are said to last many days, and long persistence is aided by the absence of other sensory experiences, as during sleep (MacKay & MacKay, 1974). Probably the decay is not a single exponential, but has a long tail, and perhaps one should not assume that there is only a single stage process. It seems premature to go into detailed discussion of the possibilities offered by hierarchies of decorrelators until the simplest mechanism has been better defined.

To summarize, we think that modifiable mutual inhibition provides a plausible explanation for pattern adaptation, and that it is a mechanism for acquiring, storing and using information about the normal pattern of contingent excitation of a group of inputs. Access to this store of knowledge occurs automatically whenever the system is used, and its effect is to make the representational elements independent in the normal environment.

4.9 Anatomical plausibility

Modern methods have enormously improved our knowledge of the neuroanatomy of the cortex, but in spite of this it is still insufficient to exclude the vast majority of models. We may, however, have succeeded in picking one of these rare ones, for we suggest (see Figure 4.5) that there are direct inhibitory connections between the cortical output neurons (mainly pyramidal cells), and such connections are not thought to exist. We have two possible answers to this. First, the inhibition may be relayed through one of the classes of neuron that are thought to mediate intra-cortical inhibition; since there is unlikely to be one such neuron for every output neuron, this would necessitate decorrelating groups of neurons rather than individual ones, which weakens some of our arguments. Another possibility is that the modifiable connections are in fact excitatory, but with the strength of the feedback modified in the anti-Hebbian direction, and perhaps superimposed on a background level of inhibition from a wider range of other cortical neurons. This is an attractive possibility since the anatomical evidence suggests that the majority of the excitatory input synapses on cortical neurons derive from the recurrent collaterals of other cortical neurons in the vicinity. At first such positive feedback sounds explosively unstable, but of course the anti-Hebbian feature would tend to stabilize it, for it would strongly discourage joint firing of many neurons. It requires more modelling to see if any such scheme might work.

Another unrealistic feature of the present model is the assumption of high accuracy real variables for input and output in place of low accuracy pulse trains. There are also problems if the number of outputs exceeds the number of degrees of freedom, or dimensionality, of the inputs. It is clearly very much an idealized version of what may actually be going on, and the attempt to make it anatomically realistic has hardly begun.

4.10 Significance for early steps in image analysis

The model works on the principle of a null instrument such as a Wheatstone bridge, where the variable element in one arm is adjusted to balance the potential drop caused by the unknown element in the other; the ease of detecting very small deviations from zero makes the method extremely sensitive and accurate. The light adaptation mechanism illustrated in Figures 4.2 & 4.3 does something of the same sort, for when the mean luminance changes, the neurons respond initially, but this is then reduced to a low level, so the neurons are again sensitive to small deviations from the new mean luminance. But in the neural network proposed here, the null principle is applied to something more interesting than a voltage or a luminance, for it is part of the associative structure of the images that is balanced out.

Now, a null instrument is particularly sensitive to whatever measure of its input is nulled; are we particularly sensitive to the local properties of images that the network adapts to and compensates for? That's too big a question to try to answer here, but it is striking how many of the primary types of visual analysis do depend upon pairwise relationships between constituent elements - what is sometimes referred to as 'second order structure'. Examples in static images are texture, orientation, and disparity, and motion can be added if one considers the analysis of successive images. The moiré effects described by Leon Glass (1969) are particularly impressive from this point of view; the method of making them produces a peak in the local auto-correlation function, and their visibility is strongly resistant to dilution by noise in the form of unpaired dots (Maloney *et al.*, 1987). Thus the detection of associative structure may be a very important part of early, local, image analysis in perception, but decorrelation may also explain certain higher level effects.

4.11 Helmholtz's unconscious inference

Helmholtz knew that the system that interprets messages from the eye has expert knowledge of the sorts of things that do and do not happen in images, and that this knowledge is applied whenever we perceive anything. The mechanism we have suggested, based upon the adaptation phenomena we described, is perhaps involved in such an interpretive system, but before showing how it might do this, some examples of inference or induction in perception will be described.

Some of the best evidence that knowledge derived from experience of the real world is used by the interpretive system comes from experiments and observations on stereoscopic vision, though of course this does not mean experience is not required in monocular vision. We shall have to describe the phenomena rather than demonstrate them, but it is much more convincing if you can see the effects for yourself; this requires a stereo-viewing system in which you can reverse the images to the two eyes, and move your head while viewing a fixed stereo-pair.

Consider first a simple pair consisting of three vertical lines seen by each eye, with the central line displaced slightly to the right in the left eye's image; the result is of course that it is seen lying slightly in front of the two flanking lines. That is normal stereoscopic vision; it is remarkable that it wasn't discovered until 1838, by Wheatstone, and its late discovery shows how unaware one is of the subtlety of perceptual mechanisms.

If you now consider interchanging left and right eye images it is easy to see that rightward displacement of the centre line in what is now the right eye's image should cause it to be seen behind the flanking pair, and that is precisely what happens. Now repeat the experiment with a stereo photograph of a complex scene such as one views in the Victorian drawing room stereoscope. (This would probably involve separating and remounting the pair.) The result is not what one would expect. At isolated points the scene with the reversed stereo-pair shows evidence of reversed apparent depth, and the whole scene creates a sense of unease, or something being wrong. Although it lacks convincing depth, the overall scene looks surprisingly normal, so that most of the evidence provided by the binocular disparities of matched features in the two images must have been rejected or ignored.

Upon reflection one sees that making use of these cues would often lead to paradoxes in perception. For instance if the scene included a table top seen from above, near and far edges would be reversed and it would not look horizontal, so the objects on it would have to be glued to it. Furthermore the occlusion of the (originally) far edge by an object would be paradoxical - the edge ought to occlude the object. It is clear that something interferes with the utilization of disparity cues that do not 'fit-in', though it is an open question at this point whether it is high-level, possibly cognitive, knowledge of facts such as the way objects rest on table tops, or local, low-level, experience of the slant implied by perspective cues and the relationships between occlusions and disparities; we would prefer the latter.

There are many perceptual phenomena that seem to imply that the brain has much stored knowledge of images and their ways, and applies this knowledge instantly and automatically. One example is the 'toytown effect', which occurs when a stereo-pair of a scene is taken with the camera positions separated by a distance greater than the separation of the two eyes. This enhances the stereo depth effect, but it also gives the impression that the scene is not real, but a reduced scale model. Other examples are the diminutions in apparent size (*micropsia*) that result from placing a prism in front of one eye that requires increased convergence, or a lens that requires additional accommodation.

Describing the perceptions resulting from unusual forms of visual stimulation Helmholtz (1925) said 'such objects are always imagined as being present in the field of vision as would have to be there in order to produce the same impression on the nervous mechanism under conditions of normal use'. A particularly striking demonstration of this rule can be obtained when looking at a projected stereo-pair. If you move your head sideways (translating, not rotating) while looking at such a scene it appears to move, following your head as you move to the left and following it back as you move to the right. If you close one eye this impression of movement instantly ceases. At one level an explanation of this effect is that stereo-disparity tells you there are objects at different distances; but if there were, and the objects were fixed, then they would be displaced relative to each other when you move your head. This doesn't happen, and the only geometrical solution is that the objects themselves moved; this is what would be happening under 'conditions of normal use', so that is what you see. Again, our perceptual mechanism seems to have expert knowledge of what happens to images in the real world; perhaps all this knowledge comes in our genes, but we think a very natural explanation follows from our model.

Normally when you move through a scene there is a predictable relation between the stereo depth information and motion parallax - the relative motions of the images of objects at different distances. Wherever in the brain motion and stereo messages come together, the pattern of covariation of the neurons sensitive to stereo and motion parallax will be such that mutual inhibition builds up between certain sets of them. Because of the special nature of paired stereo images projected on a flat surface, when you move your head you continue to get stereo depth information, but the motion parallax cue is missing; inhibition on the covariant motion parallax neurons from the stereo neurons is unopposed by the 'expected' stimulation from motion cues, and the resulting imbalance of the motion signalling neurons causes the perception of motion in the direction contrary to that expected if the scene had been real.

Of course one needs to record from neurons that behave in this way before being fully convinced, but we think it may be possible to explain many instances of perceptual inference along these lines.

Conclusions

We started by saying that, although a lot has been discovered about the patterns of impulses that occur in the visual cortex, one cannot really understand how the world is represented in our heads without knowing about the system that interprets these patterns of impulses. We think it is a promising idea that decorrelation by anti-Hebbian mutual interaction is part of this interpretive system: it would acquire, store and make allowances for the pairwise associations in past sensory experience, thereby helping us to detect the new associations that indicate new causal factors in the world around us. Obviously we should attempt to identify the components of our network with those of the anatomical network in the cortex, and we hope to find if pattern adaptation of cortical neurons follows the expectations of the theory. We also think decorrelation may play a role in determining the susceptibility of early cortical development to modified visual experience, and we hope to explore its importance in preparing a representation of sensory information that can be used efficiently by associative networks with positive Hebbian rules.

References

Barlow H.B. (1969) Pattern recognition and the responses of sensory neurones. **Ann. New York Acad. Sci. 156,** pp. 872-881

Barlow H.B., Hawken M., Kaushal T.P. & Parker A.J. (1987) Human contrast discrimination and the contrast discrimination of cortical neurons. **J. Optical Soc. America A 4,** pp. 2366-2371

Barlow H.B. & Hill R.M. (1963) Evidence for a physiological explanation of the waterfall phenomenon and figural after-effects. **Nature 200,** pp. 1345-1347

Blakemore C. (1973) The baffled brain. In: **Illusion in Nature and Art.** R.L.Gregory & E.H.Gombrich (eds.) pp. 8-47, Duckworth, London.

Blakemore C. & Campbell F.W. (1969) On the existence of neurons in the human visual system selectively sensitive to the orientation and size of retinal images. **J. Physiol. 203**, pp. 237-260

Blakemore C. & Sutton P. (1969) Size adaptation: a new aftereffect. **Science 166**, pp. 245-247

Braddick O.J., Campbell F.W. & Atkinson J. (1978) Channels in vision: Basic aspects. In: **Handbook of Sensory Physiology Vol VIII; Perception**. R. Held, H.W. Leibowicz & H.L. Teuber (eds.) pp. 1-38, Springer, New York.

Exner S. (1894) **Entwurf zu einer physiologischen erklärung der psychischen erscheinungen.** I Theil, Franz Deuticke, Leipzig und Wien.

Gibson J.J. (1933) Adaptation, after-effect and contrast in the perception of curved lines. **J. Exp. Psychol. 16,** pp. 1-31

Gibson J.J. (1937) Adaptation with negative after-effect. **Psychol. Rev. 44**, pp. 222-244

Gilinsky A. S. (1968) Orientation-specific effects of patterns of adapting light on visual acuity. **J. Optical Soc. America 58**, pp. 13-18

Glass L. (1969) Moiré effect from random dots. **Nature 223,** pp. 578-580

Helmholtz H. von (1925) **Physiological Optics.** Translated from 3rd German Edition (1910) Volume III. The theory of the perceptions of visions. Optical Society of America, Washington.

Hubel D. H. (1988) **Eye, Brain, and Vision.** Scientific American Library, New York.

Kohonen T. (1984) **Self-organization and Associative Memory.** Springer, Berlin.

Kohonen T. & Oja E. (1976) Fast adaptive formation of orthogonalizing filters and associative memory in recurrent networks of neuron-like elements. **Biol. Cybern. 21**, pp. 85-95

MacKay D.M., & MacKay V. (1974) The time course of the McCollough effect and its physiological implications. **J. Physiol. 237,** 38-39P

Maffei L., Fiorentini A. & Bisti S. (1973) Neural correlate of perceptual adaptation to gratings. **Science 182**, pp. 1036-1038

Maloney R.K., Mitchison G.J. & Barlow H.B. (1987) The limit to the detection of Glass patterns in the presence of noise. **J. Optical Soc. America A 4,** pp. 2336-2341

McCollough C. (1965) Color adaptation of edge-detectors in the human visual system. **Science 149**, pp. 1115-1116

Michael C.R. (1978) Color vision mechanisms in monkey striate cortex: simple cells with dual opponent color receptive fields. **J. Neurophysiol. 41,** pp. 1233-1249

Ohzawa I., Sclar G. & Freeman R.D. (1985) Contrast gain control in the cat's visual system. **J. Neurophysiol. 54**, pp. 651-667

Pantle A.S. & Sekuler R.W. (1968) Size detecting mechanisms in human vision. **Science 162**, pp. 1146-1148

Pfaffman C. (1941) Gustatory afferent impulses. **J. Cell. Comp. Physiol. 17,** pp. 243-258

Poggio T., Torre V. & Koch C. (1985) Computational vision and regularisation theory. **Nature 317,** pp. 314-319

Sato T. (1980) Recent advances in the physiology of taste cells. **Progress in Neurobiology 14,** pp. 25-67

Scott T.R., Yaxley S., Sienkiewicz Z.J. & Rolls E.T. (1986) Gustatory responses in the frontal opercular cortex of the alert Cynomolgous monkey. **J. Neurophysiol. 56,** pp. 876-890

Stromeyer C.F. III, (1978) Form-color after-effects in human vision. In: **Handbook of Sensory Physiology**, Vol 8. R. Held, H.W. Leibowicz & H.L. Teuber, (eds.) pp. 97-142, Springer, New York.

Vautin R.G. & Berkeley M.A. (1977) Responses of single cells in cat visual cortex to prolonged stimulus movement: neural correlates of visual after-effects. **J. Neurophysiol. 40**, pp. 1051-1065

Yamamoto T., Yugama N., Kato T. & Kawamura Y. (1985) Gustatory responses of cortical neurons in rats. II. Information processing of taste quality. **J. Neurophysiol. 53,** pp. 1356-1369