REVIEW

Unsupervised Learning

H.B. Barlow
Kenneth Craik Laboratory, Physiological Laboratory, Downing Street, Cambridge, CB2 3EG, England

What use can the brain make of the massive flow of sensory information that occurs without any associated rewards or punishments? This question is reviewed in the light of connectionist models of unsupervised learning and some older ideas, namely the *cognitive maps* and *working models* of Tolman and Craik, and the idea that redundancy is important for understanding perception (Attneave 1954), the physiology of sensory pathways (Barlow 1959), and pattern recognition (Watanabe 1960). It is argued that (1) The redundancy of sensory messages provides the knowledge incorporated in the maps or models. (2) Some of this knowledge can be obtained by observations of mean, variance, and covariance of sensory messages, and perhaps also by a method called "minimum entropy coding." (3) Such knowledge may be incorporated in a model of "what usually happens" with which incoming messages are automatically compared, enabling unexpected discrepancies to be immediately identified. (4) Knowledge of the sort incorporated into such a filter is a necessary prerequisite of ordinary learning, and a representation whose elements are independent makes it possible to form associations with logical functions of the elements, not just with the elements themselves.

1 Introduction

Much of the information that pours into our brains throughout the waking day arrives without any obvious relationship to reinforcement, and is unaccompanied by any other form of deliberate instruction. What use can be made of this impressive flow of information? In this article I hope, first, to show that it is the redundancy contained in these messages that enables the brain to build up its "cognitive maps" or "working models" of the world around it; second, to suggest initial steps by which these might be formed; and third, to propose a structure for the maps or models that automatically ensures their access and use in everyday perception, and represents percepts in a form suitable for detecting the new associations involved in ordinary learning and conditioning.

Self-organization has been a major preoccupation of those interested in neural networks since the early days, and the volume edited by Yovits

Neural Computation **1, 295–311** (1989)

et al. (1962) gives an overview of some of this work; it is interesting to compare this with the systematic and much more developed treatment in the book on the subject by Kohonen (1984). One goal has been to explain topographic projections of sensory pathways and the occurrence of feature-selective neurons without depending completely on genetic specification (see especially von der Malsburg 1973; Nass and Cooper 1975; Cooper *et al.* 1979; Perez *et al.* 1975; Fukushima 1975, 1980; Swindale 1980, 1982; Barrow 1987). Another goal has been to explain the automatic separation and classification of clustered sensory stimuli (Rosenblatt 1959, 1962; Uttley 1958, 1979). The *informon* (Uttley 1970), for example, separated frequently occurring patterns from among a background of randomly associated elements, and it mimicked many aspects of the model of Rescorla and Wagner (1972) for conditioning and learning (Uttley 1975). Grossberg (1980) mainly emphasized the interactions between supervised and unsupervised learning. The *adaptive critic* in the pole-balancing scheme described by Barto *et al.* (1983) improved learning performance by observing the pattern of recurring correction-movements and their outcomes. Self-organization may be mediated by the *competitive learning* analyzed by Rummelhart and Zipser (1985), which has been applied to the generation of feature specificity by Barrow (1987) and to a hippocampal model by Rolls (1989). The hierarchical mapping scheme of Linsker (1986, 1988) shows spontaneous self-organization, and his *infomax* principle develops further some ideas related to those Uttley (1979) proposed. Linsker's networks can produce an organization reminiscent of the cortex both spontaneously, and in response to regularities of the incoming signals. From an informational viewpoint the recent exploration by Pearlmutter and Hinton (1986) of unsupervised procedures for discovering regularities in the input is especially relevant.

Much of this paper has antecedents in the above work as well as in theories about the importance of redundancy in perception (Attneave 1954; Barlow 1959) and pattern recognition (Watanabe 1960, 1985). However, I have also tried to relate unsupervised learning to ideas about cognitive processes developed by Tolman (1932) and Craik (1943). Since these ideas provide a link with traditional psychology they will be briefly described.

1.1 Cognitive Maps and Working Models. Tolman (1932) worked within the behaviorists' tradition, but he disagreed with the rigidity of their explanations, feeling that these did not adequately convey the richness of the knowledge about their environment that maze-running rats clearly possessed and freely utilized. As he said, "behavior reeks of purpose and of cognition," and the structured knowledge of the environment that he argued for was subsequently called a *cognitive map*. Craik (1943), in his shorter, more philosophically oriented, book proposed that "thought models, or parallels, reality." These *working models* embodied the essential features and interactions in the world that fed the senses,

so that the outcomes of various possible actions could be correctly predicted; this is very similar to the way Tolman thought of cognitive maps being used by his rats.

What is the source of the extensive and well-organized knowledge of the environment implied by the possession of a cognitive map or working model? Though their structure may be genetically determined, the specific evidence they incorporate can be obtained only from the sensory messages received by the brain, and it is argued below that it is the statistical regularities in these messages that must be used for this purpose. This is an extraordinarily complex and difficult task, for it requires something like a major program of scientific research to be conducted at a precognitive level. There is plenty of room for genetic help in doing this, but once the nature of the task has been defined the statistical aspects can be approached systematically. In the next sections this is attempted for the first few steps, and a new method of finding these regularities — minimum entropy coding — is proposed.

2 Redundancy Provides Knowledge

There are genuine conceptual difficulties in applying information theory to the nervous system. These start with the paradox that although redundancy is claimed to be terribly important, sensory pathways are said to eliminate or reduce it rather than preserve it. Some of these difficulties (such as that one) disappear upon better understanding of information theory, but others do not: it is, for instance, difficult to apply the concepts when one is uncertain about the information-bearing features of the messages in nerve fibres, and about the overall plan used to represent information in the brain. In the next section these difficulties are avoided by talking about the sensory stimuli applied to the animal rather than the messages these arouse, and by doing this the definitions can be made precise.

In principle, the maximum rate of presentation of usable information to the senses can be specified if one knows the psychophysical facts about their discriminatory capacities; call this C bits/sec. Now look at the actual rate at which information is delivered, and call this H bits/sec; then the redundancy is simply $C - H$ bits/sec, or $100 \times (C - H)/C\%$. There remains a problem about measuring H, for the lower limit to its value can be calculated only if one knows all there is to know about the constraints operating in the world that gives rise to our sensations, and this point can obviously never be reached. Fortunately the concept of redundancy remains useful even if H is calculated using incomplete knowledge of the constraints, for this defines an upper limit to H and a lower limit to the redundancy.

It is confusing to refer to these $C - H$ bits/sec as information, but the technically correct term, redundancy, is almost equally misleading,

for it suggests that this part of the sensory inflow is useless or irrelevant, whereas it is the potential source of all the available knowledge about the constant or semiconstant patterns and regularities in an animal's environment. *Knowledge* is perhaps the best term for it, though it may seem paradoxical that this knowledge of the world around us can be derived only from the redundancy of the messages. The point can be illustrated by briefly considering what nonredundant sensory stimuli would be like.

Completely nonredundant stimuli are indistinguishable from random noise. Thus, such a visual stimulus would look like a television set tuned between stations, and an auditory stimulus would sound like the hiss on an unconnected telephone line. Though meaningless to the recipient, technically such signals convey information at the maximum rate because they cannot be predicted at all from other parts of the message; $H = C$ and there is no redundancy. Thus, redundancy is the part of our sensory experience that distinguishes it from noise; the knowledge it gives us about the patterns and regularities in sensory stimuli must be what drives unsupervised learning. With this in mind one can begin to classify the forms the redundancy takes and the methods of handling it.

3 Finding and Using Sensory Redundancy

Some features of sensory stimuli are almost universal. For instance, the upper part of the visual field is imaged on the lower part of the retina in an erect animal, and it is almost always more brightly illuminated. In animals such as cats that have a reflecting tapetum one usually finds that it is confined to the part receiving the image of the lower, dimmer, part of the visual field while the reflecting tapetum is replaced by a densely absorbing pigment in the part receiving the bright image; the result is to greatly reduce the amount of scattered light obscuring the image in its dimmer parts.

The many ways that redundant aspects of sensory stimuli are reflected in permanent features of the sensory system are themselves interesting, but here we are concerned with learning-like responses. To exploit redundant features the brain must determine characteristics of the stimuli that behave in a nonrandom manner, so one can consider methodically the various measures that could be made on the messages in order to characterize these regularities statistically.

3.1 Mean. One starts with the mean, taken over the recent past. In vision, this can assume any value from a few thousandths up to many thousands of cd/m^2, but it behaves in a very nonrandom manner because it tends to stay rather constant for quite long periods. I have been sitting at my desk for the past hour, and during this time the mean luminance has always been close to 10 cd/m^2; the constancy of this mean is a highly nonrandom feature and the visual system takes advantage of it to adjust

the sensitivity of the pathways to suit the limited range of retinal illuminations it will receive. Much is understood about these adaptational mechanisms, but the principles are well understood by communication engineers and I shall go ahead to consider more interesting types of redundancy. However the way in which coding by the retina changes with the mean luminance of the images is a simple paradigm of unsupervised learning, and the one we are closest to understanding physiologically.

3.2 Variance. The variance of sensory signals probably does not show the constancy over short periods combined with very large changes over long periods that is characteristic of the mean, though a walker in mist or a fish in murky water would certainly be exposed to signals with an exceptionally low range of image contrasts and hence low variance. After the transformations in the retina, taking account of changes in the variance of the input signals is actually very nearly equivalent to adjusting for the mean values of the signals in the "on" and "off" systems, and it has been suggested that such *contrast gain control* occurs in primary visual cortex (Ohzawa *et al.* 1982, 1985).

One might perhaps consider next the higher moments of the distributions of input stimuli on the many input channels, but it is hard to imagine that adapting to these would have any great advantages and I know of no evidence that natural systems respond in any way to them. Hence the next step is the large one of considering the patterns of *correlation* between the inputs on different channels.

3.3 Covariance. The simplest measure of the correlated activity of sensory pathways would be the covariance between pairs of them. Just as adaptational mechanisms take advantage of the mean by using it as an expected value and expressing values relative to it, so one might take advantage of covariance by devising a code in which the measured correlations are "expected" in the input, but removed from the output by forming a suitable set of linear combinations of the input signals. It is possible to form an uncorrelated set of signals in a neural network with a rather simple scheme of connection and rule of synaptic modification (Barlow 1989; Barlow and Földiák 1989; see also Kohonen 1984). The essential idea is that each neuron's output feeds back to the other inputs through anti-Hebbian synapses, so that correlated activity among the outputs is discouraged. Such a network would account for many perceptual phenomena hitherto explained in terms of fatigue of pattern selective elements in sensory pathways, and it also offers a mechanism for some forms of the "unconscious inference" described by von Helmholtz (1925) and modern psychologists of perception (Rock 1983). These aspects are discussed in the references cited above, and here some of the possible extensions of the principle will be mentioned.

So far it has been supposed that the covariance is worked out from paired values occurring at the same moment, but this need not be the case. Sutton and Barto (1981) have discussed temporal relationships in conditioning, and there are several synaptic mechanisms that might depend on the correlation between synaptic input at one moment and postsynaptic depolarization at a later moment; a transmitter might cause lingering "eligibility" for subsequent reinforcement, or a synaptic rewarding factor or reverse transmitter released by a depolarized neuron might be optimally picked up by presynaptic terminals some moments after they had themselves been active. Decorrelating networks based on such principles would "expect" events that occurred in often-repeated sequences, and would tend to respond less strongly to frequently occurring sequences and more strongly to abnormal ones. It is easy to see how such a mechanism might explain aftereffects of motion.

A consequence of using covariances is that, since the inputs are taken in pairs, the number of computations increases in proportion to the square of the number of inputs. This means that it would be impossible to decorrelate the whole sensory input; the best that could be done would be to decorrelate local sets of sensory fibers. However, the process could then be repeated, possibly organizing the decorrelated outputs of the first stage according to principles other than their topographical proximity, such as proximity in color space or similarity of direction of motion (Barlow 1981; Ballard 1984). Such hierarchical decorrelation processes may have considerable potential, but there is no denying that the methods so far considered only begin the task of finding regularities in the sensory input.

3.4 Rules for Combination or Agglomeration. Decorrelation separates variables that are correlated, but if the correlation between two variables is very strong they might be conveying the same message, and then one should combine them. For instance, taste information is carried by a large number of nerve fibers each of which has its characteristic mixture of sensitivities to the four primary qualities, salt, sweet, sour, and bitter. We have shown (Barlow and Földiák 1989) how these can be decorrelated in groups of four to yield the four primary qualities, but one might expect all the outputs for one quality then to be combined on to a much smaller number of elements, for without doing this they just seem to replicate the information needlessly.

There is need for an operation of this sort in many situations: for instance, to exploit the fact that there are only two dimensions of color (in addition to luminance), to exploit the prevalence of edges in ordinary images, to combine in one entity the host of sensory experiences for which we use a single word or name, and to do the same for a commonly repeated phrase or cliché. Pearlmutter and Hinton (1986) consider a related problem, that of finding input patterns that occur more often than would be expected if the constituent features occurred independently.

Finding that some combinations occur more often than expected is the converse of finding that some combinations do not occur at all, as is the case when the number of degrees of freedom or dimensions in a set of messages is less than would appear from the form of the messages. The set of N features spans less than an N-dimensional space because certain combinations do not occur, and exploiting this is just the kind of simplification that would enable one to make useful cognitive maps and working models. Principle component analysis will do what is required, and it is believed that the method described in the next section will also, but it is natural to look for network methods, especially as these have already achieved some success (for example, Oja 1982; Rumelhart and Zipser 1985; Pearlmutter and Hinton 1986; Földiák 1989).

3.5 Minimum Entropy Coding. As with decorrelation the idea is to find a set of symbols to represent sensory messages such that, in the normal environment, each symbol is as nearly as possible independent of the others, but there are two differences: first, it is applicable to discretely coded, logically distinct variables rather than continuous ones, and second it takes into account all possible nonrandom relations between the outputs, not just the pairwise relationships of the covariance matrix. To make the principle clear the simple example of coding keyboard characters in 7 binary digits to find alternatives to the familiar 7-bit ASCII code will be considered. The advantages of examining this are its familiarity, its simplicity, and the fact that samples of normal English text are readily available from which the nonrandom character frequencies can be determined.

If a sample of ordinary text is regarded simply as a string of independent characters randomly selected from the alphabet with the probabilities given by their frequency of occurrence in ordinary text, the average entropy of the characters H_c is given by the familiar expression:

$$H_c = -\sum P_i \log P_i \tag{3.1}$$

where P_i are the probabilities of the mutually exclusive set of characters.

Each of the characters is represented by a 7-bit word, and the entropies for each bit can be obtained by measuring their frequencies in a sample of text. The entropy expression for the bits takes the form:

$$H_i = -(P_i \log P_i + Q_i \log Q_i) \tag{3.2}$$

where H_i is the average entropy of the ith bit, P_i is its probability, and Q_i is $1 - P_i$.

An estimate of the average character entropy can be obtained by adding the 7-bit entropies, but it is important to realize that this can never be less, and will usually be greater, than the character entropy given by the original expression (3.1). The reason for this is the lack of independence between the values of the bits; if it were true for all the 7

bits that their values were completely independent of the other bits occurring in any combination, then the two estimates would be equal. The object is to find a code for which this is true, or as nearly true as possible, and the method of doing this is to find a code that minimizes the sum of the bit entropies — hence the name. If the minimum is reached and the bits are truly independent we call it a *factorial* code, since each bit probability or its complement is then a factor of the probability of each of the input states.

The maneuver can be looked at another way. The seven binary digits of the ASCII code can carry a maximum of 7 bits, but actually carry less when used to transmit normal text, for two reasons. First, the bit probabilities are a long way from 1/2, which would yield the maximum bit entropy; this form of redundancy is explicit and causes no trouble, for the probability of each of the 7 bits is available wherever they are transmitted and easily measured. Second, there are complicated interdependencies among the bits, so the conditional bit probabilities are not the same as the unconditional ones; this form of redundancy is troublesome, for it is not available wherever the bits are transmitted and to describe it completely one needs to know the conditional probabilities of each bit for all combinations of other bits. If both of these forms of redundancy were taken into account the information conveyed per ASCII word would of course be the same as H_1 of expression (1), i.e., about 4.3 bits, and no change of the code would alter this. However, changing the code does change the relative amounts of the two forms of redundancy, and by finding one that minimizes the sum of the bit entropies one maximizes the redundancy that results from bit probabilities deviating from 1/2. This leaves less room for redundancy from interdependencies between the bits; the troublesome form of redundancy is squeezed out by maximizing the other less troublesome form.

The minimum entropy principle should be generally applicable and clearly goes further than decorrelation, which considers only the outputs in pairs. It can also be used to compare and select from codes that change the number of channels or dimensionality of the messages. The entropy is a locally computable quantity, and by minimizing it one can increase the independence of the outputs without actually measuring the frequencies of all the possible output states, which would often be an impossible task. An accompanying article (Barlow *et al.* 1989) goes into some of the practical and theoretical problems in finding minimum entropy codes.

In this section it has been suggested that the statistical regularities of the incoming sensory messages might be measured and used to change the way they are coded or represented. It is easy to see that this would have advantages, analogous to those conferred by automatic gain control, in ensuring a compact representation within the dynamic range of the representative elements, but there may be more profound benefits attached to a representation in which the variables are independent in the environment to which the representation has been adapted. To under-

stand these one must consider the main task for which our perceptions are used, namely the detection of new associations and their utilization in ordinary learning and conditioning.

4 Ordinary Learning Requires Previous Knowledge

Over the past 20 years the work of Kamin (1969), Rescorla and Wagner (1972), Mackintosh (1974, 1983), Dickinson (1980), and others has brought about a very big change in the way theorists approach the learning problem. Whereas previously they tended to think in terms of mechanistic links whose strengths were increased or decreased according to definable laws, attention has now shifted to the computational problem that an animal solves when it learns. This started with the realization and experimental demonstration of the fact that the detection of new associations is strongly dependent on other previously and concurrently learned associations, many of which may be "silent" in that they do not themselves produce overt and obvious effects on outward behavior. As a result of this change it is at last appreciated that the brain studied in the learning laboratory is doing a profoundly difficult job: it is deducing causal links from which it can benefit in the world around it, and it does this by detecting suspicious coincidences; that is, it picks up associations that are surprising, new, or different among those that the experimenter offers it.

To detect new associations one must detect changes in the probabilities of certain events, and once this is realized an important role for unreinforced experience becomes clear: it is to find out and record the a priori probabilities, that is, the normal state of affairs, or what usually happens. Though this elementary fact does not seem to have been much emphasized by learning theorists it is obviously crucial, for how can something be recognized as new and surprising if there is no preexisting knowledge about what is old and expected?

4.1 Detecting New Associations. The basic step in learning is to detect that event C predicts U; C might be the conditional, U the unconditional stimulus of Pavlovian conditioning, or C might be a motor action and U a reinforcement in operant conditioning, or they might be successive elements in a learned sequence. Unsupervised learning can help with at least two aspects of this process: first, the separate representation of a wide range of alternative Cs, and second, the acquisition of knowledge of the normal probabilities of occurrence of these possible conditional stimuli.

It is often tacitly assumed that all alternative conditioning stimuli can be separated by the brain and their occurrences independently registered in some way, but one should not blandly ignore the whole problem of pattern recognition, and the massive interconnections we know exist

between the neurons of the brain means that the host of alternative Cs are unlikely to be completely separable unless there are specific mechanisms for ensuring that they are. The tacit assumption that the probabilities of occurrence of these stimuli, or of their cooccurrence with U, are known is equally unjustified, though it is evident that if they were not there would be no sound basis for judging that a particular C had become a good predictor of U. The logical steps necessary to detect an association between C and U will be considered in more detail to show the importance both of knowledge of their normal probabilities and of the separability of alternative conditional stimuli.

The only way to establish that C usefully predicts U is to disprove the null hypothesis that the number of occasions U follows C is no more than would be expected from chance coincidences of the two events; it is easy to see that if this null hypothesis is correct, no benefit can possibly result from using C as a predictor of U. To know the expected rate of chance coincidences one must either have measured the normal rate of the compound event (U following C) directly, or have knowledge of the normal probabilities of occurrence, P(C) and P(U); further if these probabilities are to be used it must be reasonable to assume they are independent. This prior knowledge is clearly necessary before new predictive associations can be detected reliably. Now consider the difficulties that arise if a particular C cannot be fully resolved or separated from the alternative Cs.

Failure of resolution or separation means that the registration of the occurrence of an event is contaminated by occurrences of other events. Estimates of the probabilities of occurrence of C both with and without U would be misleading if based on these contaminated counts, and their use would cause failures to detect associations that were present and the detection of spurious associations that did not exist. Thus, if counts of alternative events like C are to be used to detect causal factors, they must be adequately resolved or separated if learning is to be efficient and reliable.

4.2 Independence Is Needed for Versatile Learning. Now reconsider the two ways, measurement and calculation, of estimating the compound event probability P(U following C). Directly measuring it is adequate and plausible when one has prior expectations about the possible conditional stimuli C, especially as in either scheme one must somehow be able to detect the occurrence of this sequence when it occurs. But calculating P(U following C) from P(C) and P(U) is much more versatile, for the following reason. Measuring the rates of N coincidences such as "U following C" just gives these rates and no more, whereas knowledge of the probabilities of N independent events enables one to calculate the probability of all possible logical functions of those events, at least in principle. This gigantic increase in the number of null hypotheses whose predictions can be specified and tested gives an enormous advan-

tage to the method of calculating, rather than measuring, the expected coincidence rates. However, calculating P(U following C) from the probabilities of its constituents depends on the formation of a representation in which the constituent events can be relied on to be independent until the association that is to be detected occurs.

To summarize: to detect a suspicious coincidence that signals a new causal factor in the environment one should have access to prior knowledge of the probabilities of simpler constituent events, and these simpler events should be separately registered and independent on the null hypothesis from which one wishes to advance. It is obviously an enormously difficult and complicated task to generate such a representation, and the types of coding discussed above are only first steps; however, the versatility of subsequent learning must depend critically on how well the task has been done.

4.3 Some Other Issues. The approach taken here might be criticized on the grounds that the problem facing the brain in learning is considered in too abstract a manner, the actual mechanisms being ignored. For example, the logic of the situation requires that the numbers of occurrences and joint occurrences be somehow stored, and one might point to this as the major problem, rather than the way the numbers are used. It certainly is a major problem, but the attitude adopted here is that one is not going to get far in understanding learning without recognizing the logic of inductive inference, since this dictates what quantities actually need to be stored; it seems obvious that this problem should be looked at first.

There must be many ways in which the brain fails to perform the idealized operations required to detect new causal factors. It performs approximations and estimates, not exact calculations, but one cannot appreciate the mistakes an approximation will lead to without knowing what the exact calculation is. It is likely that many of the features of learning stem from the nature of the problem being tackled, not from the specific details of the mechanisms, and it is foolish to confuse the one with the other through failing to attend to the complexity of the task the brain appears to perform so effectively.

There is another somewhat irrelevant issue. If it was known with certainty that a predictive relation between C and U existed it would still have to be decided whether it should be acted on. This theoretically depends on whether P(U following C) is high enough for the reward obtained when U does follow C to outweigh the penalty attached to the behavior needed to reap the reward when U fails to materialize; that is a different matter from deciding whether the relation exists, and for the present it can be ignored.