INSTITUTE OF PHYSICS PUBLISHING NETWORK: COMPUTATION IN NEURAL SYSTEMS

Network: Comput. Neural Syst. **12** (2001) 241–253 www.iop.org/Journals/ne PII: S0954-898X(01)24263-8

Redundancy reduction revisited

Horace Barlow

Physiological Laboratory, Downing Site, Cambridge CB2 3EG, UK

E-mail: hbb10@cam.ac.uk

Received 31 November 2000

Abstract

Soon after Shannon defined the concept of redundancy it was suggested that it gave insight into mechanisms of sensory processing, perception, intelligence and inference. Can we now judge whether there is anything in this idea, and can we see where it should direct our thinking? This paper argues that the original hypothesis was wrong in over-emphasizing the role of compressive coding and economy in neuron numbers, but right in drawing attention to the importance of redundancy. Furthermore there is a clear direction in which it now points, namely to the overwhelming importance of probabilities and statistics in neuroscience. The brain has to decide upon actions in a competitive, chance-driven world, and to do this well it must know about and exploit the non-random probabilities and interdependences of objects and events signalled by sensory messages. These are particularly relevant for Bayesian calculations of the optimum course of action. Instead of thinking of neural representations as transformations of stimulus energies, we should regard them as approximate estimates of the probable truths of hypotheses about the current environment, for these are the quantities required by a probabilistic brain working on Bayesian principles.

1. History

The idea that the statistics of the sensory stimuli we receive from the environment are important for perception and cognition is not new, and surprisingly clear statements about it can be found before 1950 in the writings of Mach (1886), Pearson (1892), Helmholtz (1925), Craik (1943) and others. But Shannon's definition of channel capacity, information and redundancy (Shannon and Weaver 1949) was a landmark. The relations between these quantities, the probabilities of individual signals and the statistics of ensembles of signals, are not intuitively obvious, and they were a revelation to me—particularly as brought out in Shannon's wonderful paper on the redundancy of written English (Shannon 1951). I was then at an early stage in my scientific career, and since these measurable quantities were obviously important to anyone who wanted to understand sensory coding and perception, I eagerly stepped on the boat.

Fred Attneave (1954) had got there before me with his article in *Psychological Reviews*, which I heard about when I presented my ideas to a discussion group in Cambridge in the

0954-898X/01/030241+13$30.00

mid-1950s (Barlow 1961). My leading idea was the same as Attneave's: as he put it '... the human brain could not possibly utilize all the information provided by states of stimulation that were not redundant'. At about the same time Watanabe (1960) was also arguing that redundancy is important in inductive reasoning and inference. Luckily I was more interested in the relationship between redundancy and neuro-physiological mechanisms of sensation and perception, Attneave was more concerned with perceptual aspects and Watanabe with high-level inference, so we all had different things to say. But we agreed that physical stimuli from the natural environment are redundant in ways that must be important to an animal, that this will be reflected in the redundancy of sensory messages and that the coding and transformation of these messages at all levels could be adapted to this redundancy in advantageous ways.

The idea has had a rather chequered history since then. On the one hand some limitations of the original idea became clear and these will be considered below. Then, for a time, information theory dropped out of the limelight in neuroscience, and if the idea of redundancy reduction was mentioned at all it was often misunderstood. But over the past decade or so interest has been re-awakened, largely through the efforts of Laughlin (1981), Srinivasan *et al* (1982), Field (1987, 1994), Bialek *et al* (1991), Atick (1992), van Hateren (1992) and their collaborators. In addition the ideas are more readily testable, and since we now know many more details about the mechanisms of sensory processing in the brain we naturally want to understand its basis—i.e. the survival value of these mechanisms.

I have recently reviewed the history of redundancy in perception in the context of Shepard's idea about internalizing environmental regularities (Shepard 1984, 1994, Barlow 2001), and Simoncelli and Olshausen (2001) have recently reviewed the statistical properties of natural images, greatly clarifying many of the issues that are important here. Notice that there are some applications of information theory to neuroscience that have little direct bearing on the original redundancy hypothesis. These are concerned with the measurement of information transfer using entropy measures and are dealt with by Bialek *et al* (1991) and Rieke *et al* (1997).

The aim of this paper is to look at the original idea with the benefit of hindsight in order to preserve what was right, correct what was wrong and see where it leads. The conclusions are that the idea was right in drawing attention to the importance of redundancy in sensory messages because this can often lead to crucially important knowledge of the environment, but it was wrong in emphasizing the main technical use for redundancy, which is compressive coding. The idea points to the enormous importance of estimating probabilities for almost everything the brain does, from determining what is redundant to fuelling Bayesian calculations of near-optimal courses of action in a complicated world.

2. Clearing the ground

According to Shannon redundancy is what wastes channel capacity. He defined it as the difference between the entropy of the ensemble of messages actually transmitted and the maximum entropy of the ensemble that the channel could transmit. The simplest cause of this difference is unequal probability of occurrence of the elements of these messages (e.g. letters of the alphabet), but it can also arise from inequality of their joint probabilities, of from any other constraint on their occurrence. For the topic under discussion the difference is important, because inequality of the element frequencies is obvious and easy to discover, whereas other constraints may be non-obvious and very hard to discover. Unknown, non-manifest or hidden redundancy may not only be a source of important knowledge about the environment, but also, if ignored, it may lead to catastrophic errors in estimating the probabilities of hypotheses about the environment.

Redundancy and capacity were originally defined for discrete variables, such as the letters of the alphabet or binary variables, but it can be extended to continuous variables perturbed by noise. Many of the successful applications of redundancy in neuroscience have used the latter form, considering nerve messages as continuous variables, and this has both advantages and disadvantages. It is perhaps more logical to treat impulse frequency as a continuous variable, and it focuses attention on signal/noise problems, which are certainly important because of the noisiness of transduction and transmission in the brain. But the logic-like faculties of brains that lie behind higher mental functions are more interesting than the linear analysis that continuous signals link with most naturally, so my own bias is to regard the simpler concept of redundancy in discrete signals as more interesting and important. This should be taken into account in reading the discussion that follows.

To clear the ground I shall start with some unattributed one-line criticisms; the original idea will be defended where it still seems right, but matters on which it was wrong or irrelevant will be flagged.

2.1. Coding is selective, not reversible

Quite often Attneave's idea has been expressed as 'stripping away redundancy to leave the information that is biologically important'. This confuses *selective* coding, where some information is retained and some is deliberately discarded, with redundancy reduction, where no information need be lost. The point of the original idea was that you can achieve economy without losing any information at all, for a true redundancy reducing code is reversible—the input can be accurately reconstructed from the output. Of course selective coding does occur in sensory systems, so full reconstruction is not actually possible, but it is surely important to distinguish a process where information is lost irretrievably from reversible coding where it is not.

2.2. Redundancy is unnecessary information

This is another confusion, arising partly from the everyday rather than technical meaning of redundancy. It is important to realize that redundancy is not something useless that can be stripped off and ignored. An animal must identify what is redundant in its sensory messages, for this can tell it about structure and statistical regularity in its environment that are important for its survival. Some information about them can be conveyed in its genes, but sensory redundancy is the main source of knowledge from its own individual experience.

2.3. Redundancy is too vague a term to be useful

There is something in this criticism, for redundancy can take so many different forms that having a single word for it may be a little misleading. It has already been pointed out that manifest redundancy, caused by unequal probabilities of primary message or representational elements, is benign because it is so obvious and easily discovered. On the other hand hidden or latent redundancy, caused by unequal joint probabilities of higher-order combinations of elements, can lead to missed opportunities and erroneous conclusions. The complexity of redundancy is illustrated by the fact that to determine all possible forms of it one would have to measure the frequencies of all possible messages, and this would obviously not be practical except in the simplest instances. It is true that Shannon's definition makes it a measurable quantity, and this is certainly a step in the right direction, but one must raise the question, 'Is it the right measure?'. This is briefly considered again in section 4, paragraph 3.

Notice that these problems do not prove that the concept is useless; knowledge of any form of redundant regularity in the input messages is potentially useful—you do not have to know all forms of redundancy to exploit the forms of it that you do know about. It is however knowledge and recognition of the redundancy, not its reduction, that matters.

2.4. Shannon's redundancy is not appropriate in the brain

Shannon developed his theory with a very specific model in mind. He postulated an ensemble of messages to be transmitted whose statistics were fully known and unchanging. These were to be passed down a channel whose properties were also known accurately, and in the simplest case (sufficient for defining information, capacity and redundancy) there was no noise. Also he assumed that the messages could be subdivided into blocks as required, with delays in transmission until the whole of a block was available.

In the brain one usually needs a different model. We rarely know the statistics of the messages completely, and our knowledge may change. There is therefore a confusing temporal aspect to the process, for what is redundant today was not necessarily redundant yesterday. As knowledge of an environment is acquired, those features of sensory stimulation that are accurately predictable from that knowledge become redundant, but they are in a genuine sense not redundant until this knowledge has been acquired. Shannon's assumptions were of course right for the purpose of defining redundancy, but to use the concept in neuroscience we need to be more flexible.

We can also see clearly that delays in coding would be disruptive and dangerous. But the main objection is that the brain does not necessarily use redundancy for the purpose Shannon had primarily in mind, namely compression to allow the information to be passed down a channel of lower capacity. Although redundancy points to the importance of statistical regularities in the input messages, one cannot be certain that redundancy based on entropies is the correct measure unless compression is the main advantage to be derived from knowing about it, and there is room for doubt about this (see section 4, paragraph 3).

In neuroscience one must be cautious about using Shannon's formulation of the role of statistical regularities, because the brain uses information in different ways from those common in communication engineering.

2.5. Redundancy is mainly useful for error avoidance and correction

When Shannon's simplest model is made more complicated by assuming that the channel is not noise free but introduces errors, then redundancy in the input ensemble can make it possible to correct such errors. Since it is certainly true that sensory transducers and neural communication channels introduce noise, this is likely to be important in the brain, but the correction of such internally generated errors is a separate problem, and it will not be considered further here.

2.6. The redundancy of representation is not actually decreased

This is the point on which my own opinion has changed most, partly in response to criticism, partly in response to new facts that have emerged. Originally both Attneave and I strongly emphasized the economy that could be achieved by recoding sensory messages to take advantage of their redundancy, but two points have become clear since those early days. First, anatomical evidence shows that there are very many more neurons at higher levels in the brain, suggesting that redundancy does not decrease, but actually increases. Second, the obvious forms of compressed, non-redundant, representation would not be at all suitable for the kinds

of task that brains have to perform with the information represented; this is discussed in the next section.

In most mammals there are vastly more photoreceptors than there are fibres in the optic nerve, and it has been suggested that retinal coding can be viewed as redundancy reduction to compress information into a channel of reduced capacity (Srinivasan *et al* 1982, Atick 1992). This is an attractive idea, but photoreceptors are very much slower than optic nerve fibres, and at moderate and high luminance levels only a small proportion of them are operating within their dynamic ranges. It is therefore not clear that the reduction in capacity is as great as the numbers initially suggest. Also it can be argued that, whether or not compression into the optic nerve occurs, the most interesting applications of the idea are for the logic-like way information is processed at higher levels.

In the cortex it seems likely that channel capacity increases rather than decreases. The two optic nerves of humans contain axons from just over 2×10^6 retinal ganglion cells, whereas in V1 alone there are probably about 10^9 neurons. Initially I thought these facts might be reconciled with redundancy reduction by including the mean firing rate as a constraint when defining the capacity of a neuron, and then assuming that the mean firing rate of many of the neurons in V1 is extremely low. This would point to the central representation being extremely sparse, which is a view that I shall return to because it still has its attractions. However the numbers of neurons at different levels in the pathways have been more clearly established over the past 30 years, and estimates of their firing rates have become higher with the use of un-anaesthetized preparations. Although it still seems just possible that there is a large reserve of neurons in the cortex that are hardly ever active and hardly ever recorded from, or which come into use slowly during the lifetime of an animal, there is no convincing evidence that this is the case.

As a result of these developments I think one has to recognize that the information capacity of the higher representations is likely to be greater than that of the representation in the retina or optic nerve. If this is so, redundancy must increase, not decrease, because information cannot be created. The next point may, however, go a long way towards explaining what is going on here.

2.7. Compressed representations are unsuitable for the brain

The typical result of a redundancy-reducing code would be to produce a distributed representation of the sensory input with a high activity ratio, in which many neurons are active simultaneously, and with high and nearly equal frequencies. It can be shown that, for one of the operations that is most essential in order to perform brain-like tasks, such high-activity-ratio distributed representations are not only inconvenient, but also grossly inefficient from a statistical viewpoint (Gardner-Medwin and Barlow 2001).

Behind almost any interesting operation the brain performs, from detecting novelty to classical learning, lies the need to estimate the probability of occurrence of an input message, or of a class of input messages. In a distributed representation there is not in general any element that is active for an input one is interested in, and only active for that input. The absence of such an element poses a problem, for there will be no location in the brain where all the information required to count occurrences of the input of interest is brought together and is kept uncontaminated by occurrences of other inputs, and if this is not done the probability cannot be accurately determined. As an alternative, one is pretty well forced to estimate probabilities of inputs by combining measures of the probabilities of occurrence of the representational elements that are active for them, but this introduces a major source of error. In a distributed representation neurons are typically

active for inputs that one does not wish to count and include in a frequency estimate, as well as for the inputs of interest, and this is particularly the case if the activity ratio is high. Although one can compensate for the mean error introduced from such overlaps, there is no way to overcome the increases in the variances of the resulting estimates. High-activity-ratio distributed representations, which are the typical product of redundancy-reducing codes, lead to inaccurate estimates of frequencies, and the resulting statistical inefficiency would slow down learning or make it unreliable. This could obviously be disastrous for survival, so high-activity-ratio distributed representations are likely to be unsuitable for use as a basis for learning, or for any cognitive function that requires probability estimates.

To overcome this problem one needs representations with minimum overlap, that is ones with the minimum number of elements active in both of two inputs that need to be distinguished. Such overlap is perfectly allowable, on the other hand, when two inputs do not need to be distinguished, for example when an action learned for one of them is appropriately generalized to the occurrence of the other. Our quantitative estimates of the seriousness of this problem (Gardner–Medwin and Barlow 2001) used two models. In the simple one, reliable and efficient frequency estimates of R different input states required approximately the same number, R, of representational elements—vastly more than the $\log_2 R$ that are sufficient simply for unambiguous representation in a distributed representation with high activity ratio. One can do better than this in a more complex model that has modifiable interconnections (Gardner-Medwin 1976) between the representational elements, but the number is still much greater than $\log_2 R$. Accurate frequency estimation in a typical distributed representation requires very high redundancy, but this must be in a form that reduces overlap—one must minimize the number of elements active in more than one of the sensory stimuli that the brain needs to distinguish.

3. Why redundancy is still important

This example, where redundancy definitely helps the brain perform an important task, makes us re-assess the redundancy-reduction hypothesis, for it would have little remaining value if the compression apparently predicted by it does not occur, and if it would be harmful if it did! But this totally negative message is incomplete, for as the original hypothesis claimed, it is still true that discovering statistical structure in sensory messages is important. The point Attneave and I failed to appreciate is that the best way to code information depends enormously on the use that is to be made of it. As a general point this has been well recognized (see e.g. Levesque and Brachman 1987). In the current case, if you simply want to transmit information to another location, then redundancy-reducing codes economizing channel capacity are what you need. This is the aspect Shannon's formulation brings out most clearly and is also what is most important for communication engineers. But the brain is not just a communication system, and we now need to survey cases where compression is not the best way to exploit statistical structure. What I think emerges is that coding should convert hidden redundancy into a manifest, explicit, immediately recognizable form, rather than reduce it or eliminate it.

3.1. Improving S/N ratios

Knowledge of the properties of signals that are behaviourally important for an animal can be used to improve the signal/noise ratio for their detection by matching the characteristics of the

detector to those properties. As far as possible this preserves the stimulus energy and excludes other signals that would only contribute noise. This is important for birds detecting the songs of their own species, and similarly for crickets, bats and electric fish. A similar principle must be responsible for one's ability to pick out one's own name whispered at the other side of a noisy room, and one's dog can do the same.

These advantages are obtained by having detectors selective for particular patterns among the many that one receives. There is a less obvious advantage to be obtained by having the whole sensory system selective for the specific statistical properties of natural stimuli, as opposed to the class of all possible stimuli covering the same waveband. Natural images differ from white-noise images with the same waveband, as is shown by one's instant ability to distinguish examples of each, and Kersten (1987) used the ability of human subjects to fill in missing pixels in natural images to estimate how redundant they are. He obtained values around 50–75%, but the ability to achieve ratios of 10:1 or higher in image compression suggests the redundancy is even higher.

Similar improvements to sensitivity and signal/noise ratios can be gained from the knowledge of natural images obtained by principal components analysis, or other comparable methods, though this is not usually cited as the advantage of employing such methods. One can use the results in two ways: if knowledge of the presence, or amplitude, of a particular component is useful for some purpose, then you are in good position to use it for that purpose. But if it is known not to be useful you still gain, for you can remove it from the input message, leaving a residue where other types of information can be detected with improved signal/noise ratio because what you have removed no longer contributes to the noise.

This second course of action often seems to be occurring in the early stages of sensory pathways through the action of temporal adaptation and lateral inhibition, and it is often cited as an example of 'rejecting redundant information'. But this is the wrong way to look at it, for the low-temporal- and spatial-frequency information has not necessarily been rejected: in many cases it has simply been separated, and is adequately represented on other nerve fibres. Separation allows the appropriate extended spatial or temporal summation for the low frequencies, improving sensitivity for their detection, and it can also improve sensitivity for detecting patterns involving high frequencies by reducing interference from the low frequencies.

3.2. Prediction

Next to consider is prediction, which obviously promotes survival whether one is considering prey-capture, predator-avoidance or simply keeping ahead of the competition. Prediction is possible if there is spatio-temporal redundancy in the input data. One must first know that the data are redundant in containing more than the random amount of particular patterns with temporal characteristics, such as a constant trend in one direction, or a tendency to recur at a particular interval. Then if one can identify one of these patterns at an early stage, one can predict that it will be continued or that it will recur after some interval. Knowledge of the redundancy of the messages from the environment enables such a pattern to be identified in its early stages, and knowledge of its particular temporal pattern makes prediction possible.

Prediction is important on all timescales, from the millisecond range of a fly pursuing its mate to the years or centuries involved in forecasting eclipses. It must be particularly important in vision, because the early stages in photo-detection and transduction are so slow, and many early visual mechanisms may be concerned specifically with countering this slowness by prediction.

3.3. Associative learning

If reinforcement were only randomly associated with the sensory messages an animal receives, it could never learn reliably what caused the rewards and punishments that follow these messages. Thus there is a genuine sense in which all reinforcement learning is a response to statistical structure (of special kinds) in the sensory messages. This is quite an instructive way of looking at the problem, for it immediately brings home the fact that much more is required for learning than reinforcement. First one needs a representation in which those different external objects and events that can be learned about separately, are represented separately. Second one needs estimates of the probabilities of occurrence of these objects and events. Third, if the system is to learn about combinations of them, one needs to know that they are independent, or what the dependences are. Most familiar representations of sensory stimuli, such as a photograph or a tape recording, do none of this. Learning theorists often blandly assume that, when they change the external stimulus in an experiment, there will be reliable changes in the internal representation that can be used for learning, but this assumption needs more justification than it receives.

Perceptual mechanisms segregate objects from their backgrounds, classify them and identify them, and as far as we can judge at the moment the neural mechanisms for doing this employ a large fraction of available neurons in the brains of higher animals. The extent to which reinforcement influences classification is unresolved, but it is genuinely difficult to see where perception stops and learning begins—indeed it is not clear how much of the apparent difficulty of learning remains, once perception has properly prepared the ground for it. In classical learning the sensory stimuli that habitually precede a small number of innately specified reinforcements are identified and cause conditioned responses, but this is relatively simple. The task of perception is more general and difficult, for natural stimuli have to be classified according to their statistics in a way that allows the resulting items to be separately counted and have their probabilities estimated, and perhaps it must be ensured that the independence assumption is valid.

3.4. Increasing the information carried by active network elements

An animal switches the main goal of its behaviour rather infrequently, no more often than, say, once every few seconds. When it does so it needs summary representations of its environment in which evidence about important matters has been collected together. It does not need details of a large number of individually insignificant events, which is the form in which sensory messages normally arrive, and is incidentally also the form in which it would be presented after compressive coding. An example may make this clearer. At a fork in the road, both branches lead to many possible destinations each with its own associations, but an animal in a hurry needs simple signs directing it to food, safety or other opportunities, not the detailed evidence for such directions. For its current behavioural choices the brain needs representations in which detailed evidence has already been gathered together into chunks worth more than a single bit.

The idea of a very sparse representation, mentioned above, captures this notion, which is one reason why I hesitated to abandon it. Certainly information that is widely scattered over the brain in many unknown neurons cannot contribute to useful decision-making without appropriate means for collecting it together, and as we have seen compressive coding—the expected result of straightforward redundancy reduction—is positively harmful. But a very sparse representation with minimum overlap would be a different matter, provided that it retained as much as possible of the original information. The probability of a given element

being active would be low, so when it became active this would be worth more than just one bit, and it could make an important contribution towards recognizing an object or justifying a major change of goal. This was the thinking behind the suggestion (Barlow 1972) that perception is represented by a relatively small number of active 'cardinal cells', each with a selectivity intermediate between those of supposed 'pontifical neurons', and those of a typical distributed representation.

4. Displaying the redundancy as well as the message

The examples given above show that the statistical structure of sensory messages can be used in many different ways, but behind them all lies the fact that, since the data are redundant, it must be possible to represent them in a simpler form without introducing ambiguity. If the 'reduction' part of the redundancy reduction hypothesis is discarded one sees that exploiting redundancy in sensory messages links with much other work in statistics, artificial intelligence and neural networks, where the aim has been to find hidden factors that would account for the data. The technique of factor analysis has long been used to determine a small number of factors capable of accounting for variations in data such as intelligence test results, and principal components analysis is a more general technique for organizing the sources of variation in multi-dimensional data. Latent structure analysis (Henry 1983) postulates discrete 'attitudes' to account for the results of social surveys, and in the field of artificial intelligence Neal (1992) has described a network that determines simplifying beliefs, Hinton and Zemel (1994) suggest minimum description length and vector quantization methods for finding latent variables and Bishop *et al* (1998) describe a method that uses nonlinear transformations but is otherwise analogous to factor analysis. In all these methods the aim is to determine these hidden factors and make them explicit. This aim would be achieved by sensory coding that not only represents incoming messages unambiguously, but also makes explicit the ways in which they are redundant.

Notice that written language achieves something like this, for all fluent readers can use the probabilities of the components, as Shannon's analysis (1951) showed, and therefore must, in some sense, have knowledge of them. Economizing in the number of impulses used to transmit messages, rather than the number of neurons employed, would produce a coded output with its redundancy explicit in the form of non-equal frequencies of use of the primary message elements. The principle of 'economy of impulses' was actually proposed originally as a form of redundancy reduction (Barlow 1961), but it may be better to regard it as converting redundancy into a standard form that can be recognized wherever the neuron's activity is detectable.

If all the redundancy is to be in this standard form, the elements would also have to be active independently of each other in the normal environment. One would then have a factorial representation with the merit that the frequencies of conjunctions of two or more elements can be estimated immediately from the product of their individual frequencies. There is therefore no hidden redundancy; it is all manifest in the non-optimal frequencies of activity in the elements. Sparse coding of this type would obviously be exceedingly difficult to achieve with the large number of elements occurring in sensory systems, and it could only be done in some hierarchical fashion (e.g. Barlow 1981). This might explain why a large increase in the total number of neurons seems to be required, but notice that the object of such coding is to represent information in a form where redundancy is manifest, not one in which it has been reduced.

Abbott and Dayan (1999) have pointed out the curious fact that positive correlations between the representational elements can sometimes improve the accuracy of representation in a population code. Another curious fact is that negative correlations between elements in a representation would decrease the overlap and thus increase the efficiency for making frequency estimates. Negative correlations are, however, a form of statistical structure and

must increase redundancy. It is thus unclear whether Shannon's redundancy is the appropriate measure for the statistical properties of input messages that are important in the brain. What may be needed is coding for overlap-reduction, rather than for reduction of redundancy based on entropies.

Allman (1990) drew attention to the high energy cost of the mammalian neocortex, and argued that it only evolved because it increased the food-finding efficiency of its owner. A detailed energy budget shows that impulse activity in the cortex does in fact require extraordinarily high energy consumption (Laughlin and Attwell 2000), lending support to the suggestion by Levy and Baxter (1996) that economy of impulses is necessary because it reduces energy consumption. This argument for economy of impulses is of course independent of the notion that it is used to make the redundancy of sensory messages explicit, but there is no reason why both advantages should not be important.

To summarize this section, economy in the number of neurons used for the representation of sensory information is a bad idea, and the reverse is what actually seems to happen. On the other hand economy in the number of *active* neurons would make redundancy manifest and explicit rather than hidden, and would make each impulse represent an important, informative, event. We now need to step back and take a more global view of the brain's task in order to see what lies behind the importance of recognizing redundancy.

5. Statistics of natural stimuli and Bayesian inference

To determine the best (i.e. most probably rewarded, least probably punished) way for an animal to behave at any time its brain should decide what hypotheses about the world around the animal hold true at that time. The brain must therefore derive the probabilities of hypotheses being true from the evidence currently provided by its senses, and this is what Bayes' expression tells one how to do. If $P(H|D)$ is the probability of a hypothesis being true, given the current sensory data, then

$$P(H|D) = P(D|H) \times P(H)/P(D)$$

where $P(D|H)$ is the probability of the data given that the hypothesis is true, $P(H)$ is the prior probability of the hypothesis being true and $P(D)$ is the probability of the data.

All four components of Bayes' expression are probabilities. This re-emphasizes the importance of having distributed representations in which frequencies of occurrence can be reliably and accurately determined, as discussed in section 2.7, but it also brings out the importance of the statistical structure of the input messages. The first term on the rhs is the probability of obtaining the current pattern of sensory data, given that some particular hypothesis about the current environment is true. This requires a model of the way the real world causes the sensory messages that come from it, and this in turn has to be derived from the statistical structure of these sensory messages, aided of course by innately determined assumptions. To put this in a different way, Craik's working models (Craik 1943) are the bridge between the sensory messages received by an animal, and the hypothetical objects and events in its environment whose truth-probabilities the brain needs in order to decide upon its appropriate actions. These models incorporate and must be largely derived from the observed statistical regularities in the sensory stimuli, which explains the importance of sensory redundancy.

The lesson from this is that we should be thinking how perceptual mechanisms form probability estimates, how probabilities are represented in the brain and how they are transmitted from place to place in it. This is not a new idea (see Helmholtz 1925, Barlow 1969, Gregory 1970) but we defer accepting it because we persist in thinking of sensory

messages and perceptions just as transformations of the physical stimuli. It is, however, the probabilities that are required to select appropriate behaviour.

Can this insight help? Here are some brief points that do not answer the question, but make one feel optimistic about the possibility of answers.

5.1. How to modify old (inherited) probabilities with new evidence

All the variables in Bayes' expression are probabilities, but to obtain from scene statistics the three probabilities on the rhs of Bayes' expression of course requires something more. The prior probability of a hypothesis being true gives trouble to those who insist that a probability has no meaning unless it is based on measured frequencies, but this need not be a stumbling block here. A biophysical variable in a cell can have a value that is initially determined genetically, and subsequently adjusted according to what happens to that cell. The innate value can surely be regarded as a probability estimate derived from the frequencies of survival and death involved in natural selection, while the adjusted value is an improved estimate utilizing frequencies of events within the experience of a particular cell.

5.2. Models for calculating likelihoods

The second term on the rhs of Bayes' expression requires a model showing how likely the hypothesis is to generate the data, if it is true. Where can this model come from? This too could start as a neural structure evolved under natural selection, with the initial connectivity and parameter values determined genetically, but modifiable subsequently by experience. We do not yet have many clues about what such variables may be, their genetics, or how they might be modified, but the Bayesian view at least provides a framework within which these questions can be asked.

Observe that it is not only the initial parameter values and connectivity that must have evolved through natural selection, but also the means for modifying these according to experience—i.e. the algorithms for executing Bayes' rule. But any naturalist could give many other equally impressive examples of evolutionary adaptation.

6. Conclusions

I doubt if it is useful for the neuroscientist to regard perception as a compressed representation of sensory experience, for compression generally implies high-activity-ratio distributed representations, and frequency estimates can only be made from these slowly or unreliably. But the brain does need a representation with as little hidden redundancy as possible: the probabilities of occurrence of the objects and events represented in it should be obvious or easily accessed, and statistical dependences between them should either be absent, or easily obtained.

In any such representation probabilities are key elements, because they are the fuel for accurate decision making. So the take-home message for the neuroscientist should be: 'Think probabilities: What probabilities are needed? How are they represented? How are they estimated? How are they modified? How are they transmitted to other places in the brain? And how are they combined for making the moderately rational decisions that we observe brains making?'. Relating the terms in Bayes' expression to measurable cellular and physiological quantities in the brain may be a difficult task, but it does not seem an impossible one.

Acknowledgments

I have learned much of what I know about redundancy from discussions with my colleagues. In the early days these included Tommy Gold, Donald MacKay, Albert Uttley and Phillip Woodward, who were fellow members of a group called the Ratio Club that met in London in the 1950s at the National Hospital for Neurological Diseases in Queen's Square. More recently it has included most of my colleagues, but especially Roland Baddeley, Tony Gardner-Medwin, Dan Kersten, Simon Laughlin, Graeme Mitchison and Dan Ruderman. The errors are probably my own.

References

Abbott L and Dayan P 1999 The effect of correlated variability on the accuracy of a population code *Neural Comput.* **11** 91–101

Allman J 1990 The origin of the neocortex *Neurosciences* **2** 257–62

Atick J J 1992 Could information theory provide an ecological theory of sensory processing? *Network* **3** 213–51

Attneave F 1954 Informational aspects of visual perception *Psychol. Rev.* **61** 183–93

Barlow H B 1961 The coding of sensory messages *Current Problems in Animal Behaviour* ed W H Thorpe and O L Zangwill (Cambridge: Cambridge University Press) pp 331–60

——1969 Pattern recognition and the responses of sensory neurones *Ann. Acad. Sci.* **156** 872–81

——1972 Single units and sensation: a neuron doctrine for perceptual psychology? *Perception* **1** 371–94

——1981 Critical limiting factors in the design of the eye and visual cortex. The Ferrier Lecture, 1980 *Proc. R. Soc.* B **212** 1–34

——2001 The exploitation of regularities in the environment by the brain *Behav. Brain Sci.* **24**

Bialek W, Rieke F R, de Ruyter van Steveninck and Warland D 1991 Reading a neural code *Science* **252** 1854–7

Bishop C M, Svensen M and Williams C K I 1998 GTM: the generative topographic mapping *Neural Comput.* **10** 215–34

Craik K J W 1943 *The Nature of Explanation* (Cambridge: Cambridge University Press)

Field D J 1987 Relations between the statistics of natural images and the response properties of cortical cells *J. Opt. Soc. Am.* A **4** 2379–94

——1994 What is the goal of sensory coding? *Neural Comput.* **6** 559–601

Gardner-Medwin A R 1976 The recall of events through the learning of associations between their parts *Proc. R. Soc.* B **194** 375–402

Gardner-Medwin A R and Barlow H B 2001 The limits of counting accuracy in distributed neural representations *Neural Comput.* **13** 477–504

Gregory R L 1970 *The Intelligent Eye* (London: Wiedenfeld)

Helmholtz H von 1925 *Physiological Optics. Volume III. The Theory of the Perceptions of Vision* (Translated from 3rd German edn, 1910) (Washington, DC: Optical Society of America)

Henry N W 1983 Latent structure analysis *Encyclopedia of Statistical Sciences* vol 4, ed S Kotz and N L Johnson (New York: Wiley) pp 497–504

Hinton G E and Zemel R S 1994 Autoencoders, minimum description length and Helmholtz free energy *Advances in Neural Information Processing Systems* vol 6, ed J D Cowan, G Tesauro and J Alspector (San Mateo, CA: Morgan Kaufmann)

Kersten D 1987 Predictability and redundancy of natural images *J. Opt. Soc. Am.* A **4** 2395–400

Laughlin S B 1981 A simple coding procedure enhances a neuron's information capacity *Z. Naturf.* c **36** 910–2

Laughlin S B and Attwell D 2000 An energy budget for glutamergic signalling in grey matter of the rat cerebral cortex *J. Physiol.* **525P** 61P

Levesque H J and Brachman R J 1987 Expressiveness and tractability in knowledge representation and reasoning *Comput. Intell.* **3** 78–93

Levy W B and Baxter R A 1996 Energy efficient neural codes *Neural Comput.* **8** 531–43

Mach E 1886 *The Analysis of Sensations, and the Relation of the Physical to the Psychical* (Translation of the 1st, revised from the 5th, German edn by S Waterlow) (Chicago, IL: Open Court) (also reprint 1959 (New York: Dover))

Neal R M 1992 Connectionist learning of belief networks *Art. Intell.* **56** 71–113

Pearson K 1892 *The Grammar of Science* (London: Scott)

Rieke F, Warland D, Steveninck R, de R van and Bialek W 1997 *Spikes: Exploring the Neural Code* (Cambridge, MA: MIT Press)

Shannon C E 1951 Prediction and entropy of printed English *Bell Syst. Tech. J.* **30** 50–64

Shannon C E and Weaver W (ed) 1949 *The Mathematical Theory of Communication* (Urbana, IL: University of Illinois Press)

Shepard R N 1984 Ecological constraints on internal representation: resonant kinematics of perceiving, imagining, thinking and dreaming *Psychol. Rev.* **91** 417–47

——1994 Perceptual–cognitive universals as reflections of the world *Psychon. Bull. Rev.* **1** 2–28

Simoncelli E P and Ohlshausen B A 2001 Statistical properties of natural images *Annu. Rev. Neurosci.* **25**

Srinivasan M V, Laughlin S B and Dubs A 1982 Predictive coding: a fresh view of inhibition in the retina *Proc. R. Soc.* B **216** 427–59

van Hateren J H 1992 A theory of maximizing sensory information *Biol. Cyber.* **68** 23–9

Watanabe S 1960 Information-theoretical aspects of inductive and deductive inference *IBM J. Res. Dev.* **4** 208–31