Chapter 7

Information theory and sensory perception

M.D. Plumbley & S.A. Abdallah

*Department of Electronic Engineering, Queen Mary University of London,
London, UK.*

Abstract

In this chapter, we explore Shannon's information theory and what it may tell us about the biological processes of sensory perception. We begin by discussing some historical theories of sensory perception, including concepts of objects and invariances as well as principles from Gestalt psychology. We consider the ecological principle that perception should be useful for an organism, and introduce information theory as a possible way to measure this usefulness. After introducing some of the key concepts from information theory, such as entropy, information, redundancy and factorial coding, we draw parallels between communication in engineering systems and sensory perception. We discuss Attneave's early proposal that perception should create an economical description of sensory inputs, as measured by information theory, and Barlow's suggestion of lateral inhibition as a possible mechanism to achieve this. We discuss the important role played by noise in an information processing system, and its importance when determining how to optimise the communication of information. This leads to the concept of information-optimal filters which change their characteristics at different signal-to-noise levels, a feature exhibited by fly and human visual systems. For information processed across topographic maps, optimal information processing leads to a principle of uniform information density, whereby a larger area in the cortex allows higher information density, and hence higher accuracy or sensitivity, at the corresponding sensors. We also discuss some recent investigations of information theory applied to spiking neural systems, and the concept of economy of impulses. Finally, we discuss some related concepts such as structure in a perceptual stimulus and implications for Gibson's ideas about perceptual systems, and we speculate about the possible importance of attention and active perception for efficient information processing in a sensory system.

1 Introduction

The problem of sensory perception has been of interest to philosophers and psychologists for hundreds of years. During the last 50 years or so, since the introduction of Shannon's information theory [1], the parallel between perception and communication has been explored by many

researchers, such as Attneave [2] and Barlow [3]. More recently, this approach has led to the development of adaptive neural network models that optimise information [4–6], and to considerations of energy-efficient sensory coding [7].

In the first part of this chapter, we will give a brief overview of some historical theories of sensory perception. This will lead us to the consideration of perception as something which is *useful* to an organism, and therefore to the idea that perception should be performed *efficiently*. From a design viewpoint, we are therefore looking to assess how we measure the usefulness of a particular sensory system, and its efficiency in performing the task of perception. In the second part of the chapter, we will consider how concepts from *information theory*, such as entropy, mutual information, and redundancy, can be used to measure this usefulness. We will see how these help to explain certain features of sensory systems, and what features are desirable in such a system.

2 Theories of perception

2.1 What is perception?

According to Fodor ([8], p. 40), ‘what perception must do is to so represent the world as to make it accessible to thought’. Thus the central problem of perception is to construct such a representation from the collection of signals emanating from sensory transducers. Fodor goes on to say, in more precise terms, that although these signals are best thought of as representing conditions at the *surface* of an organism, the representations produced by perceptual systems are ‘most naturally interpreted as characterising the arrangements of *things in the world*’. The process of going from one to the other is essentially one of *inference*, where ‘proximal stimulus’ provides the evidence or ‘premises’, and the conclusions are ‘representations of the character and distribution of distal objects’. Similarly, Barlow [9] describes perception as ‘the computation of a representation that enables us to make reliable and versatile inferences about associations occurring in the world around us’.

As a conceptual framework for thinking about perception, this is by no means a recent development. For example, in the 18th century, Berkeley wrote ([10], § 18)

It remains therefore that if we have any knowledge at all of external things, it must be by reason, inferring their existence from what is perceiv’d by sense.

Notwithstanding his use of the word ‘perceived’, the gist is the same. Helmholtz’s theories about vision and audition, from early in the 20th century, also hold up remarkably well today. With regard to vision, he suggested that ([11], § 26)

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, the eyes being used under normal ordinary conditions.

This is an early expression of the idea that perception may be concerned with inferring the worldly *causes* of sensations, the sensations themselves being incidental. He added that ‘we are wont to disregard all those parts of the sensations that are of no importance so far as external objects are concerned’.

Addressing the commonplace observation that perception seems to be a transparent, effortless activity, Helmholtz [11] argued that this occurs through a process of ‘unconscious induction’. However, Gibson [12] claimed that ‘the senses can obtain information about objects in the world

without the intervention of an intellectual process'. By *intellectual process*, he almost certainly did not mean what we might now call a computational process. What he probably objected to is the need for perceptual *inference*, because, he maintained, under normal conditions there is enough information available to leave no room for uncertainty.

Gibson also rejected the idea that passive analysis of transduced sensory signals forms the sole basis for perception ([13], Ch. 4): 'The inputs of the nerves are supposed to be the data on which perceptual processes in the brain operate. But I make a quite different assumption'. This leads to the concept of *active perception*, where a perceptual system is engaged not just in the passive analysis of whatever stimulation happens to be playing over an organism's receptors, but in the active exploration of an 'ambient stimulus field'. Thus the visual system is responsible not just for the analysis of retinal data, but also for the control of systematic eye and head movements designed to extract more information from the available stimulus. It is this information which removes the ambiguity that would otherwise necessitate an inferential system. For example, even slight movements of the head provide strong, unambiguous cues for localisation in 3D for both vision and audition. Biosonar in bats and dolphins is another good example of an intrinsically active perceptual system.

Incidentally, Helmholtz [11] touched upon similar ideas while discussing his thesis that we perceive *causes*, arguing that 'it is only by voluntarily bringing our organs of sense in various relations to the objects that we learn to be sure as to our judgments of the causes of our sensations'. This is a view of active perception as a kind of *experimentation*, by which we distinguish causal relationships from mere coincidence.

2.2 The objects of perception

Just what is it that we actually perceive? In the context of auditory perception, Bregman ([14], p. 9) put it like this: 'The goal of scene analysis is the recovery of separate descriptions of each separate thing in the environment. What are these things?' A fairly uncontroversial answer would be that they are *objects*. Bregman goes on to discuss how objects serve as foci for all the perceptual qualities one experiences in response to a scene. This is implicit in the idea of a *property*: it must be a property of *something*. The object (as a mental construct) plays a syntactic role, binding together properties pertaining to the same physical object.

Syntactic structure aside, what is the phenomenal manifestation of *object-ness*? What aspect of the stimulus triggers the mental organisation? A common observation is that a reliable *association* or *correlation* between a number of features signals the presence of an object, as noted by Berkeley ([10], Part 1, § 1):

As several of these [sensations] are observ'd to accompany each other, they come to be marked by one name, and so to be reputed as one thing. Thus, for example, a certain colour, taste, smell, figure and consistence having been observ'd to go together, are accounted one distinct thing, signified by the name *apple*.

Mach [15], in a similar vein, thought that bodies, as we perceive them, are made up of 'complexes of sensations' with some relative permanence, and, according to Helmholtz [11], 'experience shows us how to recognise a compound aggregate of sensations as being the sign of a simple object'.

The prevailing view is that we perceive the world in terms of objects, their qualities, and their adventures. An object has coherence and continuity, but also variability. The object, as a mental construct, also serves as a scaffold on which to affix properties that seem to be 'hanging around

together' in a suspicious way. These mental objects tend to agree with external objects because the latter do indeed tend to leave a trail of physical evidence, like an animal in the woods, that sometimes leaves tracks, spoor, broken twigs, remnants of its last meal and so on. Someone who knows what to look for can infer the existence of the animal, and learn a lot about it, without directly seeing it.

Gibson [12, 13] took a different position, suggesting *invariance* as a more fundamental construct than objects. Invariants can only be discovered in the context of change, whether in a single thing evolving over time, or conceptually across a set of distinct things ([12], p. 275): 'In the case of the persisting thing, I suggest, the perceptual system simply extracts the invariant from the flowing [stimulus] array; it *resonates* to the invariant structure or is *attuned* to it. In the case of the substantially distinct things, I venture, the perceptual system must *abstract* the invariants.' In short, invariants represent things that remain constant despite change, like the shape, size, and indeed the identity of an object as it moves about.

The Gestalt psychologists were concerned with organising parts into wholes. These *gestalten*, or emergent shapes, can have properties, or *gestaltqualitäten*, that have meaning only for the whole and not its parts. Köhler [16] gives the example of a musical interval: *fifthness* is not a quality possessed by notes, but by the gestalt, the interval. *Major-ness* or *minor-ness* of chords is another good example. Gestalt theory sees perceptual organisation as a grouping process, in which parts of the visual image (and they were defined primarily in the context of vision) are allocated to one or other object. These have subsequently been adapted for use in audition [14, 17], largely via the adoption of auditory 'images' which take the form of time–frequency distributions such as spectrograms. Energy in different regions of the time–frequency plane is grouped according to such criteria as onset synchrony, proximity in time or frequency, harmonicity, and the principle of *common fate*, which recognises that sounds that undergo, for example, common frequency or amplitude modulations, are likely to have come from the same source.

2.3 Dealing with uncertainty

Sensory evidence is sometimes uncertain, inconclusive or ambiguous. Usually, we do not notice this uncertainty, filling in the gaps or making assumptions based on prior experience of what is likely, with all of this happening below the level of conscious awareness. In speech, for example, phonemes in the middle of words or phrases which have been completely masked by noise are often subjectively heard without the listener even noticing that they were missing; this is known as *phonemic restoration* [18], a specific type of *perceptual restoration*. A similar effect has been reported for music [17].

Gestalt theory deals with this type of uncertainty using the principle of good closure, or *prägnanz*. When there are competing alternative organisations suggested by the other rules, or when there is missing data, this principle says that *good forms* should be preferred. Here, the term 'good form' refers to qualities such as simplicity, regularity, symmetry, or continuity. However, as Bregman ([14], p. 26) observes, we *do* sometimes see forms with gaps in them. The principle is really for completing *evidence* which has gaps in it, and its job is to fill in missing or doubtful data.

One may then ask, what does the brain choose to fill-in the gaps with? What is 'good' form? Hochberg [19] discusses how the Gestalt principle of good *prägnanz* can be interpreted as assuming either the *simplest* (in some sense), or the *most likely* resolution of an ambiguity. This suggests

that there is an interpretation of this principle in terms of likelihood and probability. Mach ([15], § 10.8, p. 213) observed:

If the visual sense acts in conformity with the habits which it has acquired under the conditions of life of the species and the individual, we may, in the first place, assume that it proceeds according to the principle of probability; that is, those functions which have most frequently excited together before, will afterwards tend to make their appearance together when only one is excited.

Brunswick (according to Hochberg [19]) thought that the ‘Gestalt laws were merely aspects of stimulus organisation, reflecting the probability that any parts of the visual field belonged to the same object’ and Bregman ([14], p. 24) wrote ‘It seems likely that the auditory system, evolving as it has in such a world, has developed principles for “betting” on which parts of a sequence of sensory inputs have arisen from the same source.’ We will see later how these ideas can be interpreted in terms of *redundancy* of sensory information.

2.4 Representation and cognition

The acquisition of perceptual knowledge can be identified with the formation of *representations* that fulfil certain goals, or that make plain relevant or interesting aspects of that which is being represented. Fodor ([8], p. 29) states that ‘contemporary cognitive theory takes it for granted that the paradigmatic psychological process is a sequence of transformations of mental representations and that the paradigmatic cognitive system is one which effects such transformations’ and he goes on to give an informal definition of *representation* ([8], p. 38): ‘Any mechanism whose states covary with environmental ones can be thought of as registering information about the world; ... the output of such systems can reasonably be thought of as *representations* of the environmental states with which they covary.’

Fodor’s ‘paradigmatic psychological process’, a sequence of representations each of which is a transformation of the previous one, can be recast as a sequence of *black box* processing units connected by communications channels. Indeed, Marr ([20], p. 3) notes a duality between processing and representation. In doing this, our focus is shifted to the characteristics of the potentially imperfect or noisy channels. If our channels are imperfect, then the concepts of *distance* and *similarity* become important. If a representation is slightly corrupted, what will it become? If we cannot be sure that we recognise an object, what others will it ‘look like?’

Richardson [21] proposed that subjective judgements of similarity between stimuli be represented as distances between points in the Euclidean plane. This has since developed into the notion that psychological dissimilarity can be equated with the idea of a distance in some metric space. As Davidson [22] points out, this is not automatically a valid identification, as geometric distances are required to satisfy a number of constraints, such as symmetry, whereas psychological dissimilarities need not. With this caveat in mind, the geometric visualisation of similarity is still a useful and intuitive aid to understanding.

Distances and representations tend to be tied together quite closely. Distance measures operate on representations of pairs of objects: if we change the representation of our objects, the distances will change too. So, if we wanted to build an artificial cognitive system that would report similarities between objects that correspond to human subjects, one way would be to use representations such that a simple distance measure would produce corresponding distances between objects.

Shepard’s principle of *psychophysical complementarity* [23] addresses the issue of perceived distance. It says that mental structures should reflect certain transformational symmetries of the real world objects they represent. If an object can be subjected to a group of transformations

(say, rotations) without changing its essential identity, this should be mirrored by a corresponding group of transformations of the internal representation. These would not necessarily be implemented as physical rotations or in a physically circular structure, but as a set of states and operators that relate in the same way. For example, they might replicate the composition of large and small rotations, and the cyclic nature of rotation.

2.5 Mental structure vs stimulus structure

Hochberg [19] discusses the debate between those who believe that perception is driven by mental structure and those who believe that it is driven by stimulus structure. Hochberg defines stimulus structure as *intra-stimulus constraints*, and mental structure as *intra-response constraints*. Shepard [23] put it more directly: ‘Does the world appear the way it does because the world is the way it is, or because we are the way we are?’

Gestalt theory holds that perceptual organisation is an achievement of the nervous system, and that *gestalten* do not exist outside the organism. However, Köhler ([16], Ch. 5) did concede that this interpretation may have an objective value: it may ‘tell us more about the world around us’, have ‘biological value’, and ‘tends to have results which agree with the entities of the physical world’.

Mach ([15], § 1.13, p. 29) also had mentalist tendencies: ‘Bodies do not produce sensations, but complexes of elements make up bodies.’ However, he did not seem to imply that sensations are amalgamated into bodies in some ad-hoc manner, but rather that there are statistical regularities which are exploited. Gibson, with his theory of the *direct perception* of invariants, and the pick-up of information without an intervening ‘intellectual process’, believed that perception was driven by stimulus structure, not mental structure, even if this stimulus structure is only fully available to an active perceptual system.

Perception dominated by stimulus structure corresponds roughly with bottom-up or *data-driven* processing, which looks for structures in data without referring to any stored knowledge. Mental structure corresponds with top-down processes, variously known as *knowledge-driven* or, in psychology, *schema-driven* processes. The use of high-level knowledge in perception is not in doubt [24]. A more pertinent question is: How does high-level knowledge get to the top in the first place? In a *knowledge engineering* methodology, it is placed there by a human designer. In unsupervised learning, on the other hand, the acquisition of high-level knowledge is itself a data-driven process: it collects over time, percolating from the bottom up. For example, Leman and Carreras [25] contrast ‘*long-term data-driven* perceptual learning with *short-term schema-driven* recognition’. This is a sensible distinction to make when there is not enough structure in short spans of data to enable a fully data-driven approach, but there is enough structure in data accumulated over longer periods; the schema represents this accumulation of experience.

In conclusion, to quote Shepard’s answer to his own question [23],

(1) The world appears the way it does because we are the way we are; and (2) we are the way we are because we have evolved in a world that is the way it is.

Each of us has been immersed in this world since birth, and therefore we have had ample opportunity to build and adapt our mental structures to suit our environment.

2.6 An ecological perspective

Let us ask ourselves the question: ‘What *use* is perception?’ This immediately brings us to the viewpoint that perception must be *useful* for an organism. From a biological and evolutionary point

of view, the machinery of perception is an expensive burden—in humans, the brain is responsible for a large fraction of the body's energy consumption—and must therefore confer considerable advantages to the creature that owns it.

This view is more or less taken for granted. For example, Shepard [26] writes: 'The brain has been shaped by natural selection; only those organisms that were able to interpret correctly what goes on in the external world and to behave accordingly have survived to reproduce.' This amounts to the *ecological* approach to perception: one that recognises the mutual relationship between the organism and its environment.

The basic idea, like perceptual theorising in general, has a long history. Nearly 300 years ago Locke ([27], BII, Ch. IX, § 12) wrote:

Perception, I believe, is, in some degree, *in all sorts of animals*; though in some possibly the avenues provided by nature for the reception of sensations are so few, and the perceptions they are received with so obscure and dull, that it comes extremely short of the quickness and variety of sensation which is in other animals; but yet it is sufficient for, and wisely adapted to, the state and conditions of that sort of animals who are thus made ...

By the 1950s, evolutionary ideas were current. Mach ([15], § 13.3) speculated about the biological relevance of the 'sensations of tone' which 'constitute the means by which we distinguish large and small bodies when sounding, between the tread of large and small animals', and also that 'the highest tones presumably are of extreme importance for the determination of the direction from which a sound proceeds'. Of the perception of colour, he thought it ([15], § 6.2) 'essentially a sensation of favourable or unfavourable chemical conditions of life. In the process of adaptation to these conditions, colour-sensation has probably been developed and modified'.

Gibson [12, 13] mapped out a more complete ecological theory, stressing that an animal and its environment cannot be considered without one another. One of the elements of the theory is the perception of *affordances* ([13], p. 127): 'The affordances of the environment are what it *offers* to the animal, what it *provides* or *furnishes*, either for good or ill.' He argued that these affordances are apprehended in a quite direct fashion, that an intrinsic part of the perception of an object is an appreciation of what can be done with it. An apple says 'eat me', a predator says 'run away from me', a chair says 'sit on me'. As Gibson suggests in his definition, the distinction between *positive* and *negative* affordances ([13], p. 137) is an important part of the perceptual experience and of great relevance to the animal concerned. Shaw *et al.* [28] coined the term *attensity* to rate the relevance or usefulness of available information to a particular organism. Things that offer positive or negative affordances would have a high attensity and should be *attended to*, but things that offer neutral affordances have a low attensity, and presumably can be ignored.

Interestingly, this ecological approach has also been adopted in relation to music by Leman and Carreras [25]. While it may be objected that music is not a 'natural' phenomenon, this raises the interesting point that exposure to a musical tradition might constitute an *environment* of sorts, though a *cultural* rather than a natural one. To the extent that perceptual systems are shaped by experience rather than by evolution, there should be no difference between natural and cultural influences. The ontogenesis of the auditory system is likely to be as responsive to one as to the other [29].

2.7 Summary

To conclude the first part of this chapter, we have seen that there is a long and varied history of theories and viewpoints about perception. The ecological perspective proposes that perception performs a useful task for the organism, and therefore we are likely to find that organisms chosen

by natural selection are those with more *efficient* perceptual systems. Given that there is a limit to the amount of information that can be processed—or equivalently, that there are costs attached to information processing—it is reasonable to suppose that available resources are directed towards the detection and interpretation of events that have some *biological relevance*, such as the presence or absence of food, predators, potential mates and so on; in short, *useful* information.

In the next section we will see how Shannon's information theory [1] has been used to measure the *information* that might be useful to an organism, and how the question 'What is perception for?' leads to proposed answers to 'What should perception do?'

3 Information and redundancy

Information theory, first introduced in Shannon's 'Mathematical theory of communication' [1], has been used and developed extensively by communications engineers ever since its publication. Communications engineering is concerned with transmitting information from one place to another as efficiently as possible, given certain costs and constraints which are imposed on the communications system which we wish to use. For example, we may have a maximum number of *bits* (binary digits) per second that we can send down a certain binary transmission link, or we may have a radio transmitter with a limit on the maximum power level which we can use. Either of these define constraints, within which we must work.

Information theory gave communications engineering a precise meaning to the idea of *rate of information transmission*. This helped to explain how the properties of a communication channel can limit how fast information can be transmitted through it, and how to *code* signals to make most efficient use of such a channel. One of these results showed that a channel has an innate limit on its information rate, its *capacity*. It is impossible to send information through a channel faster than that channel's capacity, no matter how the information is represented or coded.

Not long after information theory first appeared, psychologists began to see potential applications in their own field [30, 31]. In particular, Attneave [2] and Barlow [32] drew at least a partial analogy between communication and sensory perception, the idea being that the initial stages of a sensory system serve to communicate information about the outside world to higher centres in the brain. This view was not universal, however. Green and Courtis [33], for example, objected to the use of information theory, on the grounds that sensory perception had no objective alphabet of symbols and no objective set of transition probabilities, as required by Shannon's formulation of *information*. They suggested that this would mean that different people might get different amounts of 'information' from the same picture or visual scene. Nevertheless, the last few years have seen an increasing interest in the use of information theory in the exploration of sensory systems, giving us some insight into why perceptual systems are organised as they are.

Before we consider these approaches in more detail, let us briefly introduce the key ideas from information theory itself.

3.1 Entropy and information

The two central concepts of information theory are those of *entropy* and *information* [1]. Generally speaking, the entropy of a set of outcomes is the uncertainty in our knowledge about which outcome will actually happen: the less sure we are about the outcome, the higher the entropy. If we know for sure what the outcome will be, the entropy will be zero. Information is gained by reducing entropy, for example, by making an observation of an outcome. Before the observation, our knowledge of the outcome is limited, so we have some uncertainty about it. However, after the

observation the entropy (uncertainty) is reduced to zero: the difference is the information gained by the observation.

Consider an experiment with N possible outcomes i , $1 \leq i \leq N$ with respective probabilities p_i . The entropy H of this system is defined by the formula

$$H = - \sum_{i=1}^N p_i \log p_i, \quad (1)$$

with $p \log p$ equal to zero in the limit $p = 0$.

For example, for a fair coin toss, with $N = 2$ and $p_1 = p_2 = 1/2$, we have

$$\begin{aligned} H &= - \left(\frac{1}{2} \log \frac{1}{2} + \frac{1}{2} \log \frac{1}{2} \right) \\ &= \log 2. \end{aligned}$$

If the logarithm is taken to base 2, this quantity is expressed in *bits*, so a fair coin toss has an entropy of 1 bit.

For any number of outcomes N , the entropy is maximised when all the probabilities are equal to $1/N$. In this case, the entropy is $H = \log N$. On the other hand, if one of the outcomes has probability 1 with all others having probability 0, then the entropy H in (1) is zero: otherwise, H is always non-zero and positive.

As we mentioned before, the information gained by an observation is the entropy before it, less the entropy after it. As an example, consider our coin toss again, and assume that we observe the outcome to be a ‘head’. We denote the state of the coin by the random variable Ω , and write the entropy of the coin toss before any observation by $H(\Omega)$ (Fig. 1(a)). If we denote the observed face by X , we write the *conditional entropy* of the coin after the observation as $H(\Omega|X = \text{‘head’})$, meaning ‘the entropy in Ω given we know that $X = \text{‘head’}$ ’ (Fig. 1(b)).

The situation if the outcome is a ‘tail’ is exactly the same. The information in the observation X about the coin state Ω is then written

$$I(\Omega, X) = H(\Omega) - H(\Omega|X) = \log 2,$$

i.e. 1 bit of information was gained by the observation.

We can also measure the entropy and information content of several variables. If we have a vector $X = (X_1, \dots, X_n)$ consisting of several variables X_i , then it is possible to show that

$$H(X) \leq H(X_1) + \dots + H(X_n), \quad (2)$$

Figure 1: Probabilities of coin state $\Omega = \omega$, (a) before and (b) after observation of a ‘head’.

or in other words, the entropy of the entire vector is no larger than the sum of the entropies of its elements. Equality holds when the set of elements are statistically independent. If the individual variables X_i have the same entropy $H(X_1) = H(X_2) = \dots = H(X_n)$, then we have that $H(X) \leq nH(X_1)$.

In theory, any transformation, or *encoding*, of X , will require at least $H(X)$ bits on average to transmit to a remote observer. If we know the probability distribution of X accurately, and can use this information to construct an appropriate code, we can get as close to this theoretical limit as we like, but we can never compress X below this limit without some loss of information.

For an example, imagine a passage of English text. In a typical computer system this is encoded in ASCII symbols, taking 8 bits per letter. Should we be able to encode the information more compactly than this? What is the entropy of English text?

As a first step, consider that (ignoring spaces, punctuation and capitalisation) we have 26 letters. If each of these had equal probability, we would have $H = \log 26 = 4.7$ bits per letter. However, this is the maximum value that H can take: for any non-uniform distribution H will be less than this. For a closer estimate of the entropy H , we could measure the probabilities $p_i, i = 1, \dots, 26$, for each different letter of the alphabet. This would take into account the statistics of individual letters, giving an estimate of about 3 bits per letter, but it would not take into account the dependencies *between* letters. For example, if we know a sentence begins with 'T', then there is quite a high probability that the next letter is 'H', and the following letter is 'E' (take a look at some of the sentences beginning with 'T' in this chapter). To get a better estimate of the entropy, Shannon [34] introduced a trick: a 'guessing game' that (in theory) performs a lossless encoding of a text passage. Measuring the code length (entropy) of this encoding gives an alternative estimate of the entropy of text, which turns out to be very low.

The idea of the guessing game is to take advantage of the language knowledge of a human subject to generate an efficient encoding for text. The guessing game works like this. Suppose that you have two identical twins, who always behave deterministically and identically. An experimenter has a passage of text, and asks the first twin to try to guess each letter, continuing to guess until the correct letter is guessed. This experimenter, who we call the *encoder*, records the number of guesses taken by the first twin to correctly guess each letter. The sequence of numbers, representing the number of guesses taken for each original letter, is passed to a second experimenter, called the *decoder*. The decoder asks the second twin to guess each letter of the passage of text, pretending to have the real text passage. However, the decoder simply asks the second twin to keep guessing until the required number of guesses has been made, writing down the appropriate letter reached each time. Since the twins behave identically, they will make the same sequence of guesses, and the decoder will then recover the original text. Therefore, the sequence of numbers is a *lossless encoding* of the original text passage, and the entropy of the entire sequence of numbers is equal to the entropy of the entire original passage [34].

Of course, identical twins such as these are hard to find in practice. Nevertheless, by playing the *encoding* part of the game only, it is possible to estimate the entropy of a text passage, since we know that we should, in theory, be able to decode the message. (If the reader is unhappy with the idea of identical twins, imagine instead a robot or computer system doing the guessing, which has linguistic knowledge similar to a human.)

The point of all this is that the sequence of numbers has a probability distribution with most of the probability mass concentrated around a small number of guesses. After the first few letters of a word or sentence, the guesser will mostly guess the correct letter first time, or at least within the first few letters, so the probability of '1 guess' is much higher than '10 guesses'. This probability distribution is much more concentrated than the probability distribution for individual letters, so the entropy per symbol is likely to be smaller. In fact, if the probabilities p_i are calculated for the

sequence of guess numbers, this gives an entropy of about 1 bit per letter position. Therefore the entropy of the original text cannot be more than about 1 bit per letter position [34].

This shows that there is a significant amount of *redundancy* in a typical passage of text. Redundancy is the difference between the information capacity used to represent a signal and the information actually present in the signal, and gives us an idea of how much we can *compress* a signal by transforming it into a smaller representation, while still retaining all of the original information. Sometimes redundancy is expressed as a ratio, rather than a difference, so that it varies between zero, when the information capacity is used to the full, and one, when the information capacity is completely wasted with no information actually communicated.

3.2 Redundancy reduction in perception

Attneave [2] suggests that perception may be the process that creates an economical description of sensory inputs, by reducing the redundancy in the stimulus. To explore this idea, he adapts Shannon's guessing game to a visual scene instead of a text passage, to see what redundancy might be present. As a concrete example, he considers a picture of an inkbottle on a desk, similar to Fig. 2. This picture, as for the original text-encoding problem, is composed of discrete symbols. A pixel can take any of three values: 'black', 'white' or 'brown' ('brown', which is the colour of a wooden desk, is shown as grey in this figure). The information content of the picture is equal to its entropy, since that is the amount of information we would gain if we reduced the entropy to zero.

Now, if a subject is asked to guess the colours in the picture, scanning along the rows from left to right, taking successive rows down the picture, we would find that most mistakes are made at the edges and corners in the picture. Applying this 'Shannon game' on the scene suggests that most information is concentrated around the edges and corners of a scene. This fitted well with psychological suggestions that these features are important, and with Hubel and Wiesel's

Figure 2: A redundant visual stimulus (after [2]).

discovery of simple cells and complex cells in the visual cortex, behaving as line detectors and edge detectors [35, 36].

3.3 Redundancy reduction and decorrelation

Soon after Attneave's [2] suggestion that the visual system should create an economical description of a scene, Barlow [3] suggested that *lateral inhibition* may be a mechanism by which this is achieved. Lateral inhibition involves inhibitory connections between cells in a parallel sensory pathway, and is a well-known feature of sensory systems [36]. Any signal which is common to many neurons in this pathway will be suppressed by the action of the lateral inhibition, while differences between inputs to neighbouring neurons, such as edges in a visual scene, would produce a significant output. In this way, lateral inhibition would produce a more economical representation of the scene, and could achieve the desired *redundancy reduction*. Thus, the information present in the retina, which was estimated to have a capacity of about 10^9 bits per second [37], would be recoded in a less redundant form as it travelled through the visual system. (Interestingly, Oldfield [38] also proposed a similar mechanism of adaptive redundancy reduction by successive stages of recoding, without explicitly referring to Shannon's information theory.)

Barlow [3] further argued that these lateral inhibitory connections should be *adaptive*, allowing an efficient code to be constructed based on the statistics of the actual sensory input data. There would be some genetically specified mechanisms of redundancy reduction, to deal with ever-present environmental regularities, but there would also be a learning mechanism to deal with those conditions and contingencies peculiar to the individual. While this would require a sensory system to *learn* its final redundancy-reducing encoding, it would have the advantage that this would not have to be specified in detail in the genetic material of that organism, and would allow an organism to adapt to the particular environment in which it finds itself.

Barlow and Földiák [6] suggested a neural network learning algorithm to perform this adaptive lateral inhibition. In their network, the inhibition v_{ij} between neurons i and j in the same layer is increased in proportion to the product of their activities. Mathematically, we can express this as $\Delta v_{ij} \propto x_i x_j$ where x_i and x_j are the activities of two neurons $i \neq j$, and Δv_{ij} is a small update to the inhibition v_{ij} . When this learning algorithm has converged, the activities of the neurons have been *decorrelated* from each other: that is, the correlation $E(x_i x_j)$ of activities of neurons i and j is zero if $i \neq j$, so they are now *uncorrelated*. In contrast to a *Hebbian* learning algorithm [39], whereby the strength of *excitation* between neurons is increased when they are active together, the Barlow and Földiák algorithm causes the *inhibition* to increase, so is called an *anti-Hebbian* learning algorithm.

3.4 Factorial coding

The logical conclusion of a process of successive stages of redundancy reduction would be a completely *non-redundant* code Y . The elements Y_i of Y would be statistically independent: this is called a *factorial* code, because the joint probability distribution for independent variables factorises into

$$P(Y = y) = \prod_{i=1}^m P(Y_i = y_i). \quad (3)$$

Harpur [40] called factorial coding the 'holy grail' of unsupervised neural network learning: this is also the goal of algorithms that perform independent component analysis [41]. It means that

all structure due to dependencies between code elements Y_i has successfully been identified, accounted for, and removed.

Barlow [9] also made an argument for factorial coding based on requirements for versatile and reliable associative learning. In order to identify an association between some stimulus y and another stimulus z , an organism must notice that y accompanies z more often than would be expected by chance if y and z occurred independently. To do this efficiently it must compare the joint probability distribution $P(y, z)$ with the product $P(y)P(z)$ of marginal probability distributions: these will be equal if y and z are independent. Thus for an association to be identified in this way, the organism must be able to estimate the marginal probabilities $P(y)$ and $P(z)$. Barlow described this in the context of a single neuron, with a number of different synapses representing the elements of Y . In this case, one can imagine that each synapse is responsible for building up a picture of $P(y_i)$ using locally available information only. If the Y_i are statistically independent, eqn (3) applies, and these marginal probabilities can then be multiplied in the cell body to yield the probability of the entire stimulus, $P(y)$. He also observed that in a dynamic environment, factorial coding means that new structure can be identified as ‘new regularities’ or ‘suspicious coincidences’ that ‘betray the presence of new causal factors in the environment’ [42].

This argument explicitly recognises the desirability of modelling the probability distribution of the stimulus. Factorial coding is useful precisely in that it facilitates a particularly simple computation, but other probabilistic models might do equally well. Building and using these models might form a large part of the operation of a perceptual system.

4 Information and noise in continuous signals

So far we have used a simple discrete representation for sensory information. This is consistent with early models of neurons, which represented information using binary states, such as *firing* or *not firing* [43]. In this discrete approach, entropy and information are identical, a sensory scene is assumed to have finite entropy, and hence it contains a finite amount of information. Redundancy reduction is then the process of encoding this finite amount of information as efficiently as possible, without losing any of the information originally present (an idea that is also closely related to the concept of *minimum message length* [44] or *minimum description length* [45]).

However, on closer inspection, the real world does not appear to be composed of such discrete values. For example, a typical table on which an ink bottle might rest is not a single uniform brown colour, but consists of a seemingly continuous range of subtle shades of brown. If we used the original discrete-variable formula (1), a continuous-valued source would have an infinite number of possible values it could take. Therefore we would end up with an infinite entropy, meaning that, in theory, a visual scene might contain an *infinite* amount of information! In fact, what limits the amount of information available to our sensory systems is *noise*.

For continuous variables, we use an alternative form of entropy

$$H = - \int_{-\infty}^{\infty} p(x) \log p(x) dx. \quad (4)$$

Equation (4) is normally finite, but no longer guaranteed to be positive, and is also dependent on the scaling of variables: scaling x by a factor of n will add $\log n$ to the entropy.

The information $I(\Omega, X) = H(\Omega) - H(\Omega|X) = H(X) - H(X|\Omega)$ derived from this continuous case *is* scale independent, however, since any scaling will add the same value to both ‘before’ entropy $H(\Omega)$, and the ‘after’ entropy $H(\Omega|X)$. The entropy $H(\Omega|X)$ is the uncertainty left in Ω once we know X , and so represents the *noise* present in the observation X (Fig. 3). For example,

Figure 3: Probabilities of a Gaussian distribution before and after a noisy observation. The ‘before’ distribution has the signal entropy $H(\Omega)$, while the ‘after’ distribution has the noise entropy $H(\Omega|X)$ for an example observation $X = 4$.

for a Gaussian signal of variance $\sigma_S^2 = S$ and noise of variance $\sigma_N^2 = N$, we can calculate that the mean information gained from an observation is

$$I = 0.5 \log(1 + S/N),$$

where S/N is the signal-to-noise power (variance) ratio. As the noise power N goes to zero, we can see that the information gained will become infinite: therefore if we could measure a continuous quantity with complete accuracy, we would gain an infinite amount of information. Consideration of noise is therefore very important when determining the information available from a continuous value.

Strictly speaking, the ‘noise’ in an observation is the uncertainty due to all factors which we are not interested in. We cannot simply measure $H(X)$ any more: we either need some information about the source Ω or the noise which causes Ω and X to differ. Often noise is considered to be random effects such as thermal noise in a sensor, but in a more general case, it may simply be an aspect of a signal that we are not interested in. For example, in a cocktail party, all except one conversation going on in the room will count as ‘noise’ (something in X not due to Ω), while the one we are interested in will count as ‘signal’ Ω . Furthermore, the ‘signal’ may not be the obvious one, perhaps if the conversation taking place to our right is more interesting than the person talking in front of us! Nevertheless, in applying information theory to sensory perception, we normally assume that we are interested equally in all information available in a sensory input, apart from those random variations that would not tell us anything about the outside world.

There are many sources of this noise that might corrupt our sensory signal, and so reduce the amount of information available. Some is in the outside world, such as the *shot noise* due to the individual photons that are carrying the information about brightness of a particular surface to a retina. Other noise is present in the sensory system itself, due, for example, to variability in

behaviour of neurons, or varying amounts of transmitter substance emitted when a nerve impulse arrives at a synapse. This then leads us to the conclusion that it will be difficult for a sensory system to transmit without loss *all* the information that arrives at its receptors. This leads to the proposal that a perceptual system should adapt itself to preserve the maximum amount of information: this is Linsker's *infomax* principle [4].

4.1 Infomax and information loss

If we consider the early parts of a perceptual system such as vision to be a system for transmitting information about the environment on to higher centres in the brain, it seems reasonable to suggest that the more of the available information which is transmitted, the more effective the system will be. Some visual systems may be optimised to extract information about very specific stimuli early on: an example might be the apparent 'bug detectors' observed in the frog retina [36]. For higher animals, however, it is more likely that early parts of the visual system should process all input information equally. Linsker's *Infomax* principle therefore suggests that a perceptual system should attempt to organise itself to maximise the rate of information transmitted through it [4].

As a concrete example, suppose we have real-valued signals with independent Gaussian noise on the inputs. (For biological sensory systems, which largely use spiking neurons to communicate information [46], we might consider these real-valued signals to be encoded in the firing rate of neurons.) In the case of a sensory system consisting of a linear neural network, Linsker [4] showed that Infomax leads to the well-known linear data reduction approach of *principal components analysis* (PCA), whereby variables are transformed to retain the components in the directions of the eigenvectors with largest corresponding eigenvalues, i.e. the directions with largest variance.

An alternative view, closely related to Infomax, is that we should try to *minimise* the *loss* in information about some original signal Ω as the sensory input is processed by the perceptual system or neural network. Although this approach is in many ways equivalent to Linsker's Infomax principle, it is useful in certain technical cases, and has some interesting properties [5]. If we denote the information loss about Ω across the system (Fig. 4) which transforms X to Y by

$$\Delta I_{\Omega}(X, Y) = I(X, \Omega) - I(Y, \Omega) \quad (5)$$

then it has the following properties [5]:

1. ΔI is positive across any function f where $Y = f(X)$;
2. ΔI is positive across any additive noise Φ where $Y = X + \Phi$ (Fig. 5(a));
3. ΔI is additive across a chain of transforms (Fig. 5(b)).

Figure 4: The original signal Ω is corrupted by irrelevant noise Φ to give the stimulus X . This is then transformed by the neural network f to give the output Y .

Figure 5: Information loss is (a) positive across additive noise, and (b) additive in series.

Figure 6: Neural network output signal Y is corrupted by noise Φ .

So, to minimise the information loss across a series of transforms, the information loss across each network should be minimised. Once information about any signal Ω has been lost, it is lost for good.

4.2 Information optimisation and whitening filters

Infomax tells us that PCA is useful when noise is present on the input of a linear sensory system. However, this is not the case when the noise is on the *output* of the transform, i.e. in the sensory system itself. To investigate this situation, we can instead use an approach formulated by Shannon for economical transmission of communication signals through a noisy channel [47].

Consider a sensory system which transmits its real-valued output signal Y through a channel where it will be corrupted by additive noise Φ (Fig. [6]).

If there were no restrictions on Y , we could simply amplify the signal until we had overcome as much of the noise as we like. However, suppose that there is a power cost

$$S_T = \int_0^B S(f)^2 df, \quad (6)$$

associated with transmitting the signal Y through the channel, where $S(f)$ is the power spectral density of the signal Y at frequency f , and B is a bandwidth limit. Then, assuming both signal and noise are Gaussian, we wish to maximise the transmitted information

$$I(\Psi, X) = \int_0^B \log \frac{S(f) + N(f)}{N(f)} df \quad (7)$$

for a given power cost S_T , where $N(f)$ is the power spectral density of the noise Φ .

Using the Lagrange multiplier technique, we attempt to maximise

$$J = I(\Psi, X) - \lambda S_T = \int_0^B \left(\log \frac{S(f) + N(f)}{N(f)} - \lambda S(f) \right) df \quad (8)$$

as a function of $S(f)$ for every $0 \leq f \leq B$. This is the case when

$$S(f) + N(f) = \text{constant}, \quad (9)$$

so if $N(f)$ is *white* noise, i.e. the power spectral density is uniform, or *flat*, the power spectral density $S(f)$ should also be flat [47]. A filter which performs this flattening is called a *whitening* filter. It is well known that a signal with flat power spectral density has an autocorrelation function $R_{yy}(\tau) = E(Y(t), Y(t + \tau))$ which is proportional to a delta function $\delta(\tau)$. In other words, the time-varying output signal $Y(t_1)$ at any time t_1 should be uncorrelated with the signal $Y(t_2)$ at any other time $t_2 \neq t_1$.

This approach leads to an equivalent condition where the signal is represented over a regular grid of units in *space* instead of *time*, such as a signal from a grid of visual receptors. In this case the signal should be transformed so that the outputs of the transform are decorrelated from each other. This decorrelation is precisely the property achieved by the Barlow and Földiák [6] algorithm that we discussed in Section 3.3.

Another way to achieve this decorrelation of outputs is to use the technique of *linear predictive coding* [48]. Consider a time-sequence of input values $x_i, x_{i-1}, \dots$ where x_i is the most recent value. We can form the least mean square (LMS) linear prediction $\hat{x}_i$ of x_i from the previous values as follows:

$$\hat{x}_i = a_1 x_{i-1} + a_2 x_{i-2} + \dots, \quad (10)$$

where the coefficients a_j are chosen to minimise the expected squared error

$$\epsilon = E (x_i - \hat{x}_i)^2. \quad (11)$$

Taking the derivative of this with respect to each coefficient a_j , the condition for this minimum is

$$E [(x_i - \hat{x}_i)x_{i-j}] = 0 \quad (12)$$

for all $j > 0$. If we take the residual $y_i = x_i - \hat{x}_i$ to be the output of our transform, the LMS linear prediction gives us

$$E [y_i x_{i-j}] = 0 \quad (13)$$

for all $j > 0$, and therefore

$$E [y_i y_k] = 0 \quad (14)$$

for all $k < i$, since

$$y_k = x_k - (a_1 x_{k-1} + a_2 x_{k-2} + \dots).$$

Thus linear predictive coding will also give us the uncorrelated outputs we need.

Srinivasan *et al.* [49] suggested that predictive coding is used in the fly's visual system to perform decorrelation. They compared measurements from the fly with theoretical results based on predictive coding of typical scenes, and found reasonably good agreement at both high and low light levels. However, they did find a slight mismatch, in that the surrounding inhibition was a little more diffuse than the theory predicted.

A possible problem with the predictive coding approach is that only the *output* noise is considered in the calculation of information: the input noise is assumed to be part of the signal. At low light levels, where the input noise is a significant proportion of the input, the noise is simply considered to change the input power spectrum, making it flatter [49]. In fact, it is possible to analyse the system for both input *and* output noise (Fig. 7). We can take a similar Lagrange multiplier approach as before, and attempt to maximise transmitted information for a fixed power cost. Omitting the details, we get the following quadratic equation to solve for this optimal filter at every frequency f [50]

$$(R_c + R_r + 1)(R_c + 1) - \frac{\gamma}{N_c} R_r = 0, \quad (15)$$

Figure 7: The signal is corrupted by both input noise before the network (or filter) and output noise after the network.

Figure 8: Optimal filter change with receptor noise level. Data in (b) adapted from [37].

where R_c is the channel signal-to-noise power spectral density ratio S_c/N_c , and R_r is the receptor signal-to-noise power spectral density ratio S_r/N_r , and γ is a Lagrange multiplier which determines the particular optimal curve to be used. This leads to a non-zero filter gain G_h whenever $R_r > [(\gamma/N_c) - 1]^{-1}$. For constant N_c (corresponding to a flat channel noise spectrum) there is therefore a certain cut-off point below which noisy input signals will be suppressed completely. A similar analysis was suggested independently by van Hateren [51].

Figure 8 compares the change in this information-optimal filter, as the input noise increases by a factor of 100 relative to the signal [50], with the effect on human visual amplitude sensitivity as the light level decreases. This approach leads to sharpening, or high-pass filtering, of a stimulus at high stimulus signal-to-noise ratios (SNRs), and ‘pooling’, or low-pass filtering, of a stimulus at low SNRs. Atick and Redlich [52] proposed a related approach, developed from Barlow’s redundancy reduction hypothesis, but taking noise into account so that it could be applied to continuous-valued signals. Through the use of realistic stimuli and model parameters, they found a remarkable quantitative match with observed spatiotemporal response of the human visual system at various light levels [53]. Their group have since used this approach to investigate sensory coding of colour in fish and primates [54], and the behaviour of neurons with *lagged* and *non-lagged* responses of neurons in the *lateral geniculate nucleus*, the ‘relay station’ between the optic nerve and the visual cortex [55].

4.3 Topographic maps

So far we have looked at sensory systems where all of the channels are considered to be similar. However, biological sensory systems are often very non-uniform, concentrating resources in some areas more than others. For example, the retina is much more sensitive to detail in the centre, the *fovea*, than in the periphery. Similarly, the fingers are more sensitive to touch than the small

of the back [36]. To explore this phenomenon, we shall the consider organisation of cortical *topographic maps*.

Topographic maps are a well-known feature of sensory systems in biological organisms. These maps are organised so that nearby parts of the sensory world are represented or processed in physically nearby locations in the brain. One notable feature of these topological maps is that they can be very non-uniform in the area given over to particular parts of the sensory input. For example, the mammalian visual cortex has a very large part given over to representation of the fovea [56], and the auditory cortex of the moustached bat has a large expanded representation around 61 Hz, corresponding to the most intense harmonic of its emitted echo location pulse [57]. The ratio of cortical area to the area of sensory input which it represents is often known as the *cortical area magnification factor*, or simply the *magnification factor*.

Several approaches have been suggested for cortical map formation based on information theory. For example, Linsker [58] generated a self-organising neural map by optimising the mutual information from an input vector to the output of a network with a single layer of output neurons, arranged in a grid. Van Hulle [59] and Holthausen and Breidbach [60] also introduced information theoretic cortical map models. These models tend to use a single input value any a given time, and typically result in a magnification factor from the sensory input to the cortical map, which is proportional to the input probability density. This leads to a topographic map which has a uniform probability density for the location the single point of activation in the cortex. However, biological cortical maps are not restricted to activity in a single neuron at once, as these probability-density models tend to be. To address this, one of us [61] considered how to optimise the *information density* across a set of parallel neurons arranged in a cortical map.

We have already seen that similar Gaussian signals need to be decorrelated to make the best use of information capacity. However, we have not yet considered the possibility that these signals might have different variances, and therefore carry different amounts of information from each other. Suppose that we have n parallel sensory signals, modelled as an n -dimensional random vector Y , where each component Y_i might have a different variance. This is then transmitted via a channel which has added noise Φ , giving a received signal at the other end of the channel of

$$Z = Y + \Phi. \quad (16)$$

For simplicity, let us assume that the signal Y and noise Φ are Gaussian with covariance matrices $C_Y = E(YY^T)$ and $C_\Phi = E(\Phi\Phi^T)$ respectively, and that Φ is small. Then the information transmitted from Y to Z is

$$I(Z, Y) = \frac{1}{2} \log \det C_Z - \frac{1}{2} \log \det C_\Phi \quad (17)$$

$$\approx \frac{1}{2} \log \det C_Y - \frac{1}{2} \log \det C_\Phi, \quad (18)$$

with a power cost to transmit the information of $S_T = \text{Tr}(C_Y)$ where $\text{Tr}(\cdot)$ is the matrix trace operator, i.e. the sum of the diagonal entries of the matrix. Now, we wish to transmit our information as efficiently as possible, so we wish to maximise $I(Z, Y)$ for a given value of S_T . Using the technique of Lagrange multipliers as before, we therefore attempt to maximise the function

$$J = I(Z, Y) - \frac{1}{2} \lambda S_T, \quad (19)$$

where λ is our Lagrange multiplier. Taking the derivative of eqn (19) with respect to C_Y and equating to zero leads to the condition [62]

$$C_Y = (1/\lambda) I_n, \quad (20)$$

where I_n is the identity matrix. Thus for most efficient transmission of information through the channel, we should ensure not only that the network outputs are uncorrelated, but that they also have *equal variance*, $\sigma_Y^2 = 1/\lambda$. Even if Y is non-Gaussian, we can still say that this condition maximises the information capacity $C(Z, Y) = \max_Y I(Z, Y)$ for any given power limit S_T , and that this capacity will be achieved if Y is Gaussian.

Consider now the distribution of information in these n neurons. At the optimum point, the output variance σ_Y^2 and noise variance σ_Φ^2 of all neurons is equal. Now, the information conveyed by a scalar Gaussian signal of signal variance S plus added noise variance N is $I = 0.5 \log(1 + S/N)$, so the information transmitted across the i th neuron is

$$I_i = \frac{1}{2} \log(1 + \sigma_Y^2/\sigma_\Phi^2), \quad (21)$$

with $\sigma_Y^2 = S_T/n$. Therefore, the information I_i must be equal across each neuron i .

Now suppose these neurons are arranged into a two-dimensional map, and that the neurons are all the same size, and hence packed into the map with uniform density of h neurons per unit area. Therefore the map will have information density of

$$I' = \frac{1}{2} h \log(1 + \sigma_Y^2/\sigma_\Phi^2) \quad (22)$$

$$= \frac{1}{2} h \log(1 + S_T/N_T) \quad (23)$$

bits per unit area, where $N_T = \text{Tr}(C_\Phi) = n\sigma_\Phi^2$ is the total noise power. The important point here is that this optimal information density is the same everywhere. This leads to the proposal of the principle of *uniform information density* [61]: that the most efficient processing of sensory information in a topographic cortical map with uniform density of neurons will be achieved when the information density across the map is uniform.

So, if an organism needs higher sensitivity for certain ranges of its sensory input than others, the most information-efficient way to do this is to *magnify* the representation of these in the cortex, in proportion to the information density required at the sensors (Fig. 9). This fits qualitatively with the observation that cortical maps tend to have a large area allocated for sensory ranges where high accuracy or sensitivity is required, with a correspondingly smaller area where less sensitivity is required. Specifically we found that the principle of uniform information density was consistent with changes in behavioural threshold and corresponding cortical areas observed by Recanzone *et al.* [63] for owl monkeys trained on a tactile frequency discrimination task [61].

4.4 Energy efficiency and spiking neurons

It is well known that communication between neurons using spikes is widespread in biological sensory systems. These neurons communicate by firing one or more identical spikes along their axons, rather than having the type of continuous output response that we have considered up to now. As we mentioned earlier, we could simply consider the firing rate to encode a (positive) continuous variable, and use information optimisation principles developed for continuous variables. However, recently it has become increasingly clear that a single spike may encode a considerable amount of information on its own [64].

Spikes might also be useful for another reason. For the early stages of a sensory system, such as the visual system, we are likely to have an *information bottleneck*, where we have a limited capacity channel, such as the optic nerve, through which all visual information has to be squeezed.

Figure 9: Magnification of input range over some regions of input u , allowing a non-uniform information density $I(u)$ over u , while we have a uniform map information density $I(v)$ over v .

It therefore seems likely that the information should be encoded as compactly as possible in order for as much information as possible to be passed on to higher processing centers. However, once the information reaches the neocortex, a huge expansion in representation is possible. This indicates that a compact representation may no longer be necessary: instead we may wish for an *energy efficient* coding, reducing the energy used by the brain to store and process the information. For example, Field [65] suggested that sensory information should be represented using *sparse* codes, in which most units are inactive, instead of compact codes: Baddeley [66] suggests that these sparse codes are essentially codes with a minimal firing rate cost, and are therefore energy-efficient. Verdú [67] considered the problem of optimising information channel capacity per unit cost, finding that the maximum can be achieved with a *binary* alphabet consisting of two symbols. One symbol should be zero cost, with the other being the lowest-cost symbol per bit of information transmitted. Thus if spikes have evolved to be the cheapest symbol available to a neuron, it may be that communication using spike-like symbols is the most energy-efficient method of encoding information.

Inspired by Barlow's [68] concept of an 'economy of impulses' in the brain, Levy and Baxter [7] considered the optimisation of the ratio of information capacity to energy, finding that the optimum ratio depended on the ratio r of energy cost for the cell to emit a spike, against that for no spike. Using values of r suggested by glucose utilisation and oxygen consumption measured in the rat, they calculated optimal firing frequencies close to those found in the rat subiculum (in the limbic system). This suggests that the brain may have evolved to be energy efficient in its handling of information, rather than simply maximising information capacity.

Balasubramanian *et al.* [69] consider two regimes for information processing in an organism: *immediate* activities such as catching prey or avoiding predators, and normal *exploratory* activity. They suggest that immediate activities require maximisation of information given the average energy available, whereas the exploratory regime requires maximisation of information per unit energy cost. They found different optimal operating conditions for the organism, depending on

the regime being used. Hence the organism concentrates its information processing capacity at critical points in time, lowering this capacity (and hence the rate of energy consumption) when this full capacity is not needed.

In an early paper considering the information transmitted by spiking neurons, MacKay and McCulloch [70] argued it is more efficient for a spiking neural system to use the time delays *between* spike firings to encode information than it is to use the presence or absence of a spike in a certain time slot. They estimated that, depending on the timing accuracy, interval modulation such as this is at least nine times more efficient per impulse. Nevertheless, a largely prevailing view in psychology was that individual spikes contained little information by themselves, and that many spikes would be needed for any useful information to be transmitted. However, using a method based on signal reconstruction, Bialek *et al.* [64] showed that a small number of individual spikes were able to reconstruct a typical sensory signal quite accurately, and that each spike contained a significant amount of information about the sensory input. They also showed that information flow in spiking neurons can be very efficient, in comparison with the theoretical limits of this type of information transmission [71]. (See also the book by Rieke *et al.* [46].)

Expanding our view to encompass sets of spiking neurons, correlations between the firing times of neurons may be important to represent sensory information [72]. Recently Deco and Schurmann [73] proposed that a system of spiking neurons should be tuned so that discrimination between different input stimuli can be made as reliable as possible as quickly as possible, a principle that they call *minimal time-maximum reliability*. Using this principle, they found that clusters of synchronised neurons emerged as representations of different stimuli, very reminiscent of the *cell assemblies* proposed by Hebb [39] many years earlier. Digging down even deeper into the mechanisms of neural information transmission, looking at the synapses that connect one neuron to another, Manwani and Koch [74] found that, while single synapses provided poor information transmission on their own, a small number of redundant synapses, as found in the typical set of axonal connections from one neuron to another, would produce robust and reliable information transmission.

Many of these more recent models make quite specific assumptions, and have different models of the biology built into their system model. Nevertheless, they all share the key issues of noise and unreliability of transmission. It will be interesting to see whether these approaches can be used to form a unified theory of biological information transmission in the not-too-distant future.

5 Discussion

5.1 Redundancy and structure

The information-theoretic idea of redundancy links back to the idea of *structure* in a perceptual stimulus that was discussed in Section 2.5. Let us model a stimulus as an n -tuple of random variables $X = (X_1, \dots, X_n)$, where the individual components X_n can represent simultaneous observations from many receptors, or successive observations from a single receptor, or both. Hochberg's *intra-stimulus constraints* [19] imply that the observations are confined to some lower-dimensional subspace or manifold of the full space of X . In a stochastic system, this hard constraint can be replaced with a soft one: the distribution of X might not strictly occupy a lower-dimensional subspace, but it will have some sort of *shape*, and certain sorts of shape will imply statistical dependencies between the components of X . Informally then, one may characterise structure as any departure from purely 'random' or 'unstructured' statistical independence.

So, if X is unstructured, it will be impossible to predict one component from the others; that is, the conditional probabilities $P(X_i | \{X_j: j \neq i\})$ will be equal to the marginals $P(X_i)$. Conversely, structuredness implies that it is possible to make some of those predictions, or what Attneave [2] called ‘better-than-chance inferences’. Some components will carry information about others, so there will be redundancy in the data. In the case of images, we have seen that structure implies that some regions of an image can be roughly predicted from others. This is also true for symmetric or periodic data, and accords with Barlow’s suggestion [42] that ‘structure is anything that is regular, repeating, or symmetric’.

The principle that redundancy constitutes structure agrees with many examples of what we might intuitively identify as *structure*. In images, common patterns or features, as well as being distinctive to the eye, will be a source of redundancy because the parts of the feature tend to accompany one-another: the presence of one half is a strong indicator that the other half is to be found adjacently. Objects are also a source of redundancy, since their parts always go together. Temporal coherences of many kinds, such as that found in the overall brightness of visual scenes, in the overall loudness of auditory scenes, and in the set of probable notes in tonal music, are all further forms of redundancy.

As well as the benefit of producing a more concise encoding of sensory data, the process of identifying and removing ‘structuredness’ requires that the structure be ‘understood’ in some sense. It involves an implicit probabilistic model of the data, because in an optimal code, the encoding of a message x with probability $P(x)$ is $\log_2 P(x)$ bits long. Thus, in order to optimally determine how much resource to allocate to representing the message, one must know the probability of its occurrence.

5.2 Gibson and information

The concept of *redundancy* sheds some light on Gibson’s ideas about perceptual systems. One of the points he makes is that perceptual systems should not be defined in terms of sense *organs*, (that is, anatomically) but rather that different, possibly overlapping, sets of receptors work together to form perceptual systems [12]. Each perceptual system is dedicated to knowing a different aspect of the world: they are outward looking systems rather than inward looking ones. In the current context, we might say that perceptual systems are defined by statistical dependencies between receptors. For example, the vestibular and visual organs together constitute a system for the perception of orientation, because they both respond in a correlated fashion to changes of orientation. Similarly, olfactory receptors contribute both to the perception of smells and flavours, depending on whether the signals are correlated with events at the tip of the nose, such as sniffing actions, or with tactile sensations in the mouth, chewing actions, and so on. Therefore we should be interested not just in redundancy between successive symbols in the same channels, or between symbols in neighbouring channels, but also redundancy between different sensory modalities.

However, Gibson himself expresses concern about the use of information theory, as did others (see also Section 3). While he described perception as a process of ‘information pick-up’, he felt that the use of Shannon’s mathematical framework was inappropriate [12]. He rejected the idea that ‘information’ is that which brings about a reduction in uncertainty, falling back on a dictionary definition: ‘that which is got by word or picture’, which is perhaps rather circular in the current context. He also rejected the idea that sense data can be considered as signals in a communications system ([13], p. 63), ‘for these signals must be in code and therefore have to be decoded; signals are messages, and messages have to be interpreted’. In fact, he was quite dismissive about the ‘vast literature nowadays of speculation about the media of communication’ which he accused of being ‘undisciplined and vague’.

Responding to these concerns, Barlow [42] advocates the adoption of an information theoretic approach, while at the same time acknowledging that Shannon's original exposition of the theory as a theory of *communication*, with a transmitter, a channel, and a receiver, 'is in some ways a poor analogy for the perceptual brain, partly because we must rid ourselves of the idea that there is a homunculus to receive the messages'. Even if the terminology may seem a little strained when applied to perception, and the use of the term *information* in the technical sense might not always agree precisely with our common-sense expectations, there is a case to be made that different parts of the brain are truly in 'communication' with each other, and the mathematical forms fit very well.

5.3 Noise and irrelevant information

We saw that the consideration of noise is critical to the information theoretic approach for continuous variables. Is it reasonable to assume that we know enough about the signal and noise that we are dealing with, to be able to do this?

Marr's 'fundamental hypothesis' [20] is that a perceptual system will discard information that does not impair its ability to detect 'stable features', which he defines as those which tend to co-occur. This can be imagined to be a sort of voting system amongst receptors. If only one unit reports some stimulus, it is likely to be ignored on the basis that it is probably an error, or an aberration peculiar to the internal workings of that unit. On the other hand, if many units show concerted activity, it is more likely that this is due to events in the outside world.

Marr's hypothesis amounts to an assumption about the structure of noise, namely, that noise affects each receptor independently of the others, while real signals affect many receptors together. Consequently, this hypothesis determines which features in the input are going to be important: those which are robust to such noise. It also corresponds with the idea of redundancy: stable features are encoded redundantly, whereas noise is not. Attneave [2] also suggests that uncorrelated sensory signals tend to be interpreted as noise. He observes that even though uncorrelated noise has a high entropy, and thus could potentially carry a lot of information, both visual and auditory white noise have a dull, uniform, and uninformative texture. Redlich [75] makes a similar observation: 'Desirable input information is often encoded redundantly, so redundancy can be used to distinguish true signal from noise.'

Thus we have empirical observations that noise *tends* to appear to be non-redundant and *structureless*, implying that any structuring will not be noise and therefore will be relevant. This is quite a strong statement which somehow goes against the ecological principles outlined in Section 2.6. Relevance, or attensity, is ultimately something which cannot be decided without considering the use to which the information will be put. Atick and Redlich [52] comment that 'neither *noise* nor *signal* is a universal concept, but depend on what is useful visual information to a particular organism in a particular environment'. This *usefulness* is unlikely to be determined from the sensory input data alone. Although much of the work on information theory discussed so far concentrates on bottom-up processing, organisms are likely to have evolved top-down mechanisms, such as *attention*, that allow them to extract this useful information while ignoring information that is not so useful. Nevertheless, the best that we can do for the moment appears to be to assume that information about anything with structure might be useful, so a perceptual system should aim to keep it.

5.4 Uniform information, attention, and active perception

The principle of uniform information density for cortical maps suggests that the most efficient way to hand information is if it is spread evenly across the cortical surface. The same is likely

to hold for time: the most efficient way to process information is to take it in evenly, at a steady rate. In an exploratory system, a visual system with a fovea can be used to scan over a new scene, gradually transmitting the information found like the flying-spot scanner of a television camera. If, instead, the visual system had a retina which had the acuity of the fovea everywhere, not only would a huge number of additional receptors and optic nerve fibres be needed, but we would be confronted with a huge amount of new information on first seeing a scene, with this huge information capacity being subsequently wasted. Only when something moved in the scene would more information be available, and organisms have typically evolved to move their fovea to attend to movement, ready to extract the new information available.

However, the eye is not scanned randomly or regularly over a scene. Perhaps it is 'looking' to extract from a scene the most information (or the most information relevant to some task) in the shortest possible time, even if this process may be subconscious. We can take this even farther if we physically approach an object, and turn it over to see a side which would otherwise be hidden: clearly a process of active perception [13]. This aspect is currently missing from the information theory approaches that we have discussed in this chapter. They tend to assume that information about everything in our observation X (except random noise on the receptors) is equally important, whereas for this type of active perception we would need some sort of top-down information to tell us what in X is likely to give us information about the Ω that we wish to know about, and what can be ignored.

A few researchers have begun to use information theory to investigate the efficiency of *motor output* [76]. This approach may help to explain features of skills learning or muscle operation [77]. The use of Shannon's information to investigate active perception may need to bind these two approaches together, into a theory that considers perception much more as a complete action/observation loop, with information flowing both into and out of an organism, rather than simply a process of passive observation.

6 Conclusion

People have been trying to understand sensory perception for hundreds, if not thousands, of years. Traditionally, theories of perception have been expressed in terms of objects, qualities, form, invariants, and so on. Gestalt theory, for example, proposed that perception selects interpretations with good *prägnanz*, or form, if several interpretations are possible, filling in missing information in the process.

In contrast, the ecological perspective asks the following question: What is perception *for*? This leads us to the proposal that perception must be *useful* for something, requiring us to find a way to measure that 'usefulness'. From a design viewpoint, if we have a measure of the usefulness of perception, we can start to consider what we consider perception *should* do, in order to achieve the goal of being useful.

The measure of usefulness explored here is Shannon's *information*. Optimisation of information leads to various alternative principles that a perceptual system should adhere to, including redundancy reduction, maximisation of transmitted information (Infomax), factorial coding, uniform information density, or energy efficiency, depending on the circumstances. These principles have been applied to binary and discrete-valued models, real-valued models, and more realistic models of spiking neural sensory systems. While this may seem a large range of alternative principles, they are in fact different aspects of the same underlying principle of information optimisation, albeit with different costs and constraints.

There are still many challenges remaining to be faced in the application of information theory to sensory perception. For example, more needs to be known about the implications of this approach for spiking systems. In addition, the application of information theory to active perception, where the organism is not considered to be simply a passive observer of the world, seems yet to be explored significantly, if at all. So, while information theory has already provided many insights into the design of perceptual sensory systems, there is still a long way to go before we can say that we really understand sensory perception.

Acknowledgements

S.A. was supported by a Research Studentship from the UK Engineering and Physical Sciences Research Council (EPSRC). Part of this work was supported by EPSRC research grant GR/R54620.

References

- [1] Shannon, C.E., A mathematical theory of communication. *Bell System Technical Journal*, **27**, pp. 379–423, 623–656, 1948.
- [2] Attneave, F., Some informational aspects of visual perception. *Psychological Review*, **61**, pp. 183–193, 1954.
- [3] Barlow, H.B., Possible principles underlying the transformation of sensory messages. *Sensory Communication*, ed. W.A. Rosenblith, MIT Press: Cambridge, MA, pp. 217–234, 1961.
- [4] Linsker, R., Self-organization in a perceptual network. *IEEE Computer*, **21**(3), pp. 105–117, 1988.
- [5] Plumbley, M.D. & Fallside, F., An information-theoretic approach to unsupervised connectionist models. *Proceedings of the 1988 Connectionist Models Summer School*, eds. D. Touretzky, G. Hinton & T. Sejnowski, Morgan-Kaufmann: San Mateo, CA, pp. 239–245, 1988.
- [6] Barlow, H.B. & Földiák, P., Adaptation and decorrelation in the cortex. *The Computing Neuron*, eds. R. Durbin, C. Miall & G. Mitchison, Addison-Wesley: Wokingham, England, pp. 54–72, 1989.
- [7] Levy, W.B. & Baxter, R.A., Energy efficient neural codes. *Neural Computation*, **8**, pp. 531–543, 1996.
- [8] Fodor, J., *The Modularity of Mind*, MIT Press: Cambridge, MA, 1983.
- [9] Barlow, H.B., Conditions for versatile learning, Helmholtz's unconscious inference, and the task of perception. *Vision Research*, **30**, pp. 1561–1571, 1990.
- [10] Berkeley, G., *A Treatise Concerning the Principles of Human Knowledge*, Brown and Sons, 1937/1734.
- [11] Southall, J.P.C. (ed.), *Helmholtz's Treatise on Physiological Optics*, Dover Publications: New York, 1962. (Translation of the 1910 original.)
- [12] Gibson, J.J., *The Senses Considered as Perceptual Systems*, Houghton Mifflin: Boston, MA, 1966.
- [13] Gibson, J.J., *The Ecological Approach to Visual Perception*, Houghton Mifflin: Boston, MA, 1979.
- [14] Bregman, A.S., *Auditory Scene Analysis*, MIT Press: Cambridge, MA, 1990.

- [15] Mach, E., *The Analysis of Sensations*, Dover Publications: New York, 1959. (Translated by C.M. Williams, 1st German edn, 1886.)
- [16] Köhler, W., *Gestalt Psychology*, Liveright: New York, 1947.
- [17] Deutsch, D., Grouping mechanisms in music (Chapter 9). *The Psychology of Music*, 2nd edn, ed. D. Deutsch, Academic Press: San Diego, CA, 1999.
- [18] Warren, R.M., Perceptual restoration of missing speech sounds. *Science*, **167**, 1970.
- [19] Hochberg, J., Levels of perceptual organization (Chapter 9). *Perceptual Organization*, eds. M. Kubovy & J.R. Pomerantz, Erlbaum: Hillsdale, NJ, 1981.
- [20] Marr, D., *Vision*, Freeman: San Francisco, CA, 1982.
- [21] Richardson, M.W., Multidimensional psychophysics. *Psychological Bulletin*, **35**, pp. 659–660, 1938.
- [22] Davidson, M.L., *Multidimensional Scaling*, John Wiley & Sons: New York, NY, 1983.
- [23] Shepard, R.N., Psychological complementarity (Chapter 10). *Perceptual Organization*, eds. M. Kubovy & J.R. Pomerantz, Erlbaum: Hillsdale, NJ, 1981.
- [24] Knill, D.C. & Richards, W. (eds.), *Perception as Bayesian Inference*, Cambridge University Press: Cambridge, UK, 1996.
- [25] Leman, M. & Carreras, F., The self-organization of stable perceptual maps in a realistic musical environment. *Proc. 3emes Journées d'Informatique Musicale (3rd Computer Music Conference)*, JIM96, Caen, 1996.
- [26] Shepard, R.N., Cognitive psychology and music (Chapter 3). *Music, Cognition, and Computerised Sound*, ed. P.R. Cook, MIT Press: Cambridge, MA, pp. 21–35, 1999.
- [27] Locke, J., *An Essay Concerning Human Understanding*, Dent and Sons, 1961. (First published 1706.)
- [28] Shaw, R.E., McIntyre, M. & Mace, W.M., The role of symmetry in event perception. *Perception: Essays in Honour of James J. Gibson*, eds. R.B. MacLeod & H. Pick, Cornell University Press: Ithica, NY, 1974.
- [29] Windsor, W.L., *A Perceptual Approach to the Description and Analysis of Acousmatic Music*, PhD thesis, City University, London, 1995.
- [30] Miller, G.A. & Frick, F.C., Statistical behaviouristics and sequences of responses. *Psychological Review*, **56**, pp. 311–324, 1949.
- [31] Miller, G.A., What is information measurement? *American Psychologist*, **8**, pp. 3–11, 1953.
- [32] Barlow, H.B., Sensory mechanisms, the reduction of redundancy, and intelligence. *Proceedings of a Symposium on the Mechanisation of Thought Processes*, National Physical Laboratory, Teddington, Her Majesty's Stationery Office: London, Vol. 2, pp. 537–559, 1959.
- [33] Green, R.T. & Courtis, M.C., Information theory and figure perception: The metaphor that failed. *Acta Psychologica*, **25**, pp. 12–36, 1966.
- [34] Shannon, C.E., Prediction and entropy of printed English. *Bell System Technical Journal*, **30**, pp. 50–64, 1951.
- [35] Hubel, D.H. & Wiesel, T.N., Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology*, **148**, pp. 574–591, 1959.
- [36] Kuffler, S.W., Nicholls, J.G. & Martin, A.R., *From Neuron to Brain: A Cellular Approach to the Function of the Nervous System*, 2nd edn, Sinauer Associates Inc.: Sunderland, MA, 1984.
- [37] Kelly, D.H., Information capacity of a single retinal channel. *IRE Transactions on Information Theory*, **IT-8**, pp. 221–226, 1962.
- [38] Oldfield, R.C., Memory mechanisms and the theory of schemata. *British Journal of Psychology*, **45**, pp. 14–23, 1954.

- [39] Hebb, D.O., *The Organization of Behavior*, Wiley: New York, 1949.
- [40] Harpur, G.F., *Low Entropy Coding with Unsupervised Neural Networks*, PhD thesis, Cambridge University Engineering Department, 1997.
- [41] Hyvärinen, A., Karhunen, J. & Oja, E., *Independent Component Analysis*, John Wiley & Sons: New York, 2001.
- [42] Barlow, H.B., Banishing the homonculus. *Perception as Bayesian Inference*, eds. D.C. Knill & W. Richards, Cambridge University Press: Cambridge, UK, p. 425, 1996.
- [43] McCulloch, W.S. & Pitts, W., A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, **5**, pp. 115–133, 1943.
- [44] Wallace, C.S. & Boulton, D.M., An information measure for classification. *The Computer Journal*, **11**(2), pp. 185–194, 1968.
- [45] Rissanen, J., Modeling by shortest data description. *Automatica*, **14**, pp. 465–471, 1978.
- [46] Rieke, F., Warland, D., de Ruyter van Steveninck, R. & Bialek, W., *Neural Coding*, MIT Press: Cambridge, MA, 1996.
- [47] Shannon, C.E., Communication in the presence of noise. *Proceedings of the IRE*, **37**, pp. 10–21, 1949.
- [48] Papoulis, A., *Probability, Random Variables and Stochastic Processes*, 2nd edn, McGraw-Hill: New York, 1984.
- [49] Srinivasan, M.V., Laughlin, S.B. & Dubs, A., Predictive coding; a fresh view of inhibition in the retina. *Proceedings of the Royal Society of London, Series B*, **216**, pp. 427–459, 1982.
- [50] Plumbley, M.D. & Fallside, F., The effect of receptor signal-to-noise levels on optimal filtering in a sensory system. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-91*, Vol. 4, pp. 2321–2324, 1991.
- [51] van Hateren, J.H., A theory of maximizing sensory information. *Biological Cybernetics*, **68**, pp. 23–29, 1992.
- [52] Atick, J.J. & Redlich, A.N., Towards a theory of early visual processing. *Neural Computation*, **2**, pp. 308–320, 1990.
- [53] Atick, J.J., Could information theory provide an ecological theory of sensory processing? *Network: Computation in Neural Systems*, **3**(2), pp. 213–251, 1992.
- [54] Atick, J.J., Li, Z.P. & Redlich, A.N., Understanding retinal color coding from first principles. *Neural Computation*, **4**, pp. 559–572, 1992.
- [55] Dong, D.W. & Atick, J.J., Temporal decorrelation: a theory of lagged and nonlagged responses in the lateral geniculate nucleus. *Network: Computation in Neural Systems*, **6**, pp. 159–179, 1995.
- [56] Rovamo, J., Virsu, V. & Näsänen, R., Cortical magnification factor predicts the photopic contrast sensitivity of peripheral vision. *Nature*, **271**, pp. 54–56, 1978.
- [57] Suga, N., Cortical computational maps for auditory imaging. *Neural Networks*, **3**, pp. 3–21, 1990.
- [58] Linsker, R., How to generate ordered maps by maximising the mutual information between input and output signals. *Neural Computing*, **1**, pp. 402–411, 1989.
- [59] Van Hulle, M.M., Topology-preserving map formation achieved with a purely local unsupervised competitive learning rule. *Neural Networks*, **10**, pp. 431–446, 1997.
- [60] Holthausen, K. & Breidbach, O., Self-organized feature maps and information theory. *Network: Computation in Neural Systems*, **8**, pp. 215–227, 1997.
- [61] Plumbley, M.D., Do cortical maps adapt to optimize information density? *Network: Computation in Neural Systems*, **10**, pp. 41–58, 1999.

- [62] Plumbley, M.D., Efficient information transfer and anti-Hebbian neural networks. *Neural Networks*, **6**, pp. 823–833, 1993.
- [63] Recanzone, G.H., Jenkins, W.M., Hradek, G.T. & Merzenich, M.M., Progressive improvement in discriminative abilities in adult owl monkeys performing a tactile frequency discrimination task. *Journal of Neurophysiology*, **67**, pp. 1015–1030, 1992.
- [64] Bialek, W., Rieke, F., de Ruyter van Steveninck, R. & Warland, D., Reading a neural code. *Science*, **252**, pp. 1854–1857, 1991.
- [65] Field, D.J., What is the goal of sensory coding? *Neural Computation*, **6**, pp. 559–601, 1994.
- [66] Baddeley, R., An efficient code in V1? *Nature*, **381**, pp. 560–561, 1996.
- [67] Verdú, S., On channel capacity per unit cost. *IEEE Transactions on Information Theory*, **36**, pp. 1019–1030, 1990.
- [68] Barlow, H.B., Trigger features, adaptation and economy of impulses. *Information Processing in the Nervous System*, ed. K.N. Leibovic, Springer-Verlag: New York, pp. 209–226, 1969.
- [69] Balasubramanian, V., Kimber, D. & Berry II, M.J., Metabolically efficient information processing. *Neural Computation*, **13**, pp. 799–815, 2001.
- [70] MacKay, D.M. & McCulloch, W.S., The limiting information capacity of a neuronal link. *Bulletin of Mathematical Biophysics*, **14**, pp. 127–135, 1952.
- [71] Bialek, W., DeWeese, M., Rieke, F. & Warland, D., Bits and brains: information flow in the nervous system. *Physica A*, **200**, pp. 581–593, 1993.
- [72] Panzeri, S., Schultz, S.R., Treves, A. & Rolls, E.T., Correlations and the encoding of information in the nervous system. *Proceedings of the Royal Society Series B: Biological Sciences*, **266**, pp. 1001–1012, 1999.
- [73] Deco, G. & Schürmann, B., Spatiotemporal coding in the cortex: Information flow-based learning in spiking neural networks. *Neural Computation*, **11**, pp. 919–934, 1999.
- [74] Manwani, A. & Koch, C., Detecting and estimating signals over noisy and unreliable synapses: Information-theoretic analysis. *Neural Computation*, **13**, pp. 1–33, 2000.
- [75] Redlich, A.N., Redundancy reduction as a strategy for unsupervised learning. *Neural Computation*, **5**, pp. 289–304, 1993.
- [76] Haken, H., Information compression in biological systems. *Biological Cybernetics*, **56**, pp. 11–17, 1987.
- [77] Senn, W., Wyler, K., Clamann, H.P., Kleinle, J., Lüscher, H.R. & Müller, L., Size principle and information theory. *Biological Cybernetics*, **76**, pp. 11–22, 1997.