INFORMATION THEORY AND THE BRAIN

Edited by

ROLAND BADDELEY
University of Sussex

PETER HANCOCK
University of Stirling

PETER FÖLDIÁK
University of St. Andrews

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK www.cup.cam.ac.uk
40 West 20th Street, New York, NY 10011-4211, USA www.cup.org
10 Stamford Road, Oakleigh, Melbourne 3166, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain

First published 1999

Printed in the United States of America

Typeface Times Roman 10.25/12.5 pt. *System* 3B2 [KW]

*A catalogue record for this book is available from
the British Library*

Library of Congress Cataloguing in Publication Data
Information theory and the brain / edited by Roland Baddeley, Peter
Hancock, Peter Földiák
p. cm.
1. Neural networks (Neurobiology) 2. Neural networks (Computer
science). 3. Information theory in biology. I. Baddeley, Roland,
1965- . II. Hancock, Peter J. B., 1958- . III. Földiák, Peter,
1963- .
OP363.3.I54 1999 98-32172
612.8′2—dc21 CIP

ISBN 0 521 63197 1 hardback

Contents

1

Introductory Information Theory and the Brain

ROLAND BADDELEY

1.1 Introduction

Learning and using a new technique always takes time. Even if the question initially seems very straightforward, inevitably technicalities rudely intrude. Therefore before a researcher decides to use the methods information theory provides, it is worth finding out if these set of tools are appropriate for the task in hand.

In this chapter I will therefore provide only a few important formulae and no rigorous mathematical proofs (Cover and Thomas (1991) is excellent in this respect). Neither will I provide simple "how to" recipes (for the psychologist, even after nearly 40 years, Attneave (1959) is still a good introduction). Instead, it is hoped to provide a non-mathematical introduction to the basic concepts and, using examples from the literature, show the kind of questions information theory can be used to address. If, after reading this and the following chapters, the reader decides that the methods are inappropriate, he will have saved time. If, on the other hand, the methods seem potentially useful, it is hoped that this chapter provides a simplistic overview that will alleviate the growing pains.

1.2 What Is Information Theory?

Information theory was invented by Claude Shannon and introduced in his classic book *The Mathematical Theory of Communication* (Shannon and Weaver, 1949). What then is information theory? To quote three previous authors in historical order:

Information Theory and the Brain, edited by Roland Baddeley, Peter Hancock, and Peter Földiák.

The "amount of information" is exactly the same concept that we talked about for years under the name "variance". [Miller, 1956]

The technical meaning of "information" is not radically different from the everyday meaning; it is merely more precise. [Attneave, 1959]

The *mutual information* $I(X; Y)$ is the relative entropy between the joint distribution and the product distribution $p(x)p(y)$, i.e.,

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} \log \frac{p(x, y)}{p(x)p(y)}$$

[Cover and Thomas, 1991]

Information theory is about measuring things, in particular, how much measuring one thing tells us about another thing that we did not know before. The approach information theory makes to measuring information is to first define a measure of how uncertain we are of the state of the world. We then measure how less uncertain we are of the state of the world after we have made some measurement (e.g. observing the output of a neuron; asking a question; listening to someone speak). The difference between our uncertainty before and the uncertainty after making a measurement we then define as the amount of information that measurement gives us. As can be seen, this approach depends critically on our approach to measuring uncertainty, and for this information theory uses *entropy*. To make our description more concrete, the concepts of entropy, and later information, will be illustrated using a rather artificial scenario: one person has randomly flipped to a page of this book, and another has to use yes/no questions (I said it was artificial) to work out some aspect of the page in question (for instance the page number or the author of the chapter).

Entropy

The first important aspect to quantify is how "uncertain" we are about the input we have before we measure it. There is much less to communicate about the page numbers in a two-page pamphlet than in the *Encyclopedia Britannica* and, as the measure of this initial uncertainty, entropy measures how many yes/no questions would be required on average to guess the state of the world. Given that all pages are equally likely, the number of yes/no questions required to guess the page flipped to in a two-page pamphlet would be 1, and hence this would have an entropy (uncertainty) of 1 bit. For a 1024 (2^{10}) page book, 10 yes/no questions are required on average and the entropy would be 10 bits. For a one-page book, you would not even need to ask a question, so it would have 0 bits of entropy. As well as the number of questions required to guess a signal, the entropy also measures the smallest possible size that the information could be compressed to.

The simplest situation and one encountered in many experiments is where all possible states of the world are equally likely (in our case, the "page flipper" flips to all pages with equal probability). In this case no compression is possible and the entropy (H) is equal to:

$$H = \log_2 N \tag{1.1}$$

where N is the number of possible states of the world, and $\log_2$ means that the logarithm is to the base 2.[1] Simply put, the more pages in a book, the more yes/no questions required to identify the page and the higher the entropy. But rather than work in a measuring system based on "number of pages", we work with logarithms. The reason for this is simply that in many cases we will be dealing with multiple events. If the "page flipper" flips twice, the number of possible combinations of word pages would be $N \times N$ (the numbers of states multiply). If instead we use logarithms, then the entropy of two-page flips will simply be the sum of the individual entropies (if the number of states multiply, their logarithms add). Addition is simpler than multiplication so by working with logs, we make subsequent calculations much simpler (we also make the numbers much more manageable; an entropy of 25 bits is more memorable than a system of 33,554,432 states).

When all states of the world are not equally likely, then compression is possible and fewer questions need (on average) to be asked to identify an input. People often are biased page flippers, flipping more often to the middle pages. A clever compression algorithm, or a wise asker of questions can use this information to take, on average, fewer questions to identify the given page. One of the main results of information theory is that given knowledge of the probability of all events, the minimum number of questions on average required to identify a given event (and smallest that the thing can be compressed) is given by:

$$H(X) = \sum p(x) \log_2 \frac{1}{p(x)} \tag{1.2}$$

where $p(x)$ is the probability of event x. If all events are equally likely, this reduces to equation 1.1. In all cases the value of equation 1.2 will always be equal to (if all states are equally likely), or less than (if the probabilities are not equal) the entropy as calculated using equation 1.1. This leads us to call a distribution where all states are equally likely a maximum entropy distribution, a property we will come back to later in Section 1.5.

[1] Logarithms to the base 2 are often used since this makes the "number of yes/no" interpretation possible. Sometimes, for mathematical convenience, natural logarithms are used and the resulting measurements are then expressed in nats. The conversion is simple with 1 bit $= \log(e)/\log(2)$ nats ≈ 0.69314718 nats.

Information

So entropy is intuitively a measure of (the logarithm of) the number of states the world could be in. If, after measuring the world, this uncertainty is decreased (it can never be increased), then the amount of decrease tells us how much we have learned. Therefore, the information is defined as the difference between the uncertainty before and after making a measurement. Using the probability theory notation of $P(X|Y)$ to indicate the probability of X given knowledge of Y (conditional on), the mutual information ($I(X;\,Y)$) between a measurement X and the input Y can be defined as:

$$I(X;\,Y) = H(X) - H(X|Y) \tag{1.3}$$

With a bit of mathematical manipulation, we can also get the following definitions where $H(X,\,Y)$ is the entropy of all combination of inputs and outputs (the joint distribution):

$$I(X;\,Y) = \begin{cases} H(X) - H(X|Y) & \text{(a)} \\ H(Y) - H(Y|X) & \text{(b)} \\ H(X) + H(Y) - H(X,\,Y) & \text{(c)} \end{cases} \tag{1.4}$$

1.3 Why Is This Interesting?

In the previous section, we have informally defined information but left unanswered the question of why information theory would be of any use in studying brain function. A number of reasons have inspired its use including:

Information Theory Can Be Used as a Statistical Tool. There are a number of cases where information-theoretic tools are useful simply for the statistical description or modelling of data. As a simple measure of association of two variables, the mutual information or a near relative (Good, 1961; Press et al., 1992) can be applied to both categorical and continuous signals and produces a number that is on the same scale for both. While correlation is useful for continuous variables (and if the variables are Gaussian, will produce very similar results), it is not directly applicable to categorical data. While χ^2 is applicable to categorical data, all continuous data needs to be binned. In these cases, information theory provides a well founded and general measure of relatedness.

The use of information theory in statistics also provides a basis for the tools of (non-linear) regression and prediction. Traditionally regression methods minimise the sum-squared error. If instead we minimise the (cross) entropy, this is both general (it can be applied to both categorical and continuous outputs), and if used as an objective for neural networks, maximising information (or minimising some related term) can result in neural network learning algorithms that are much simpler; theoretically more elegant; and in many cases appear to perform better (Ackley et al., 1985; Bishop, 1995).

Analysis of Informational Bottlenecks. While many problems are, for theoretical and practical reasons, not amenable to analysis using information theory, there are cases where a lot of information has to be communicated but the nature of the communication itself places strong constraints on transmission rates. The time-varying membrane potential (a rich informational source) has to be communicated using only a stream of spikes. A similar argument applies to synapses, and to retinal ganglion cells communicating the incoming light pattern to the cortex and beyond. The rate of speech production places a strong limit on the rate of communication between two people who at least sometimes think faster than they can speak. Even though a system may not be best thought of as simply a communication system, and all information transmitted may not be used, calculating transmitted information places constraints on the relationship between two systems. Looking at models that maximise information transmission may provide insight into the operation of such systems (Atick, 1992a; Linsker, 1992; Baddeley et al., 1997).

1.4 Practical Use of Information Theory

The previous section briefly outlined why, in principle, information theory might be useful. That still leaves the very important practical question of how one could measure it. Even in the original Shannon and Weaver book (Shannon and Weaver, 1949), a number of methods were used. To give a feel for how mutual information and entropy can be estimated, this section will describe a number of different methods that have been applied to problems in brain function.

Directly Measuring Discrete Probability Distributions

The most direct and simply understood method of measuring entropy and mutual information is to directly estimate the appropriate probability distributions (P(input), P(output) and P(input and output)). This is conceptually straightforward and, given enough data, a reasonable method.

One example of an application where this method is applicable was inspired by the observation that people are very bad at random number generation. People try and make sequences "more random" than real random numbers by avoiding repeats of the same digit; they also, under time pressure, repeat sequences. This ability to generate random sequences has therefore been used as a measure of cognitive load (Figure 1.1), where entropy has been used as the measure of randomness (Baddeley, 1956). The simplest estimators were based on simple letter probabilities and in this case it is very possible to directly estimate the distribution (we only have 26 probabilities to estimate). Unfortunately, methods based on simple probability estimation will prove unreliable when used to estimate, say, letter pair probabilities (a statistic that will be sensitive to some order information). In this case there are 676 (26^2) probabilities to be estimated, and subjects' patience would probably be exhausted before enough data had been collected to reliably estimate them. Note that even when estimating 26 probabilities, entropy will be systematically underestimated (and information overestimated) if we only have small amounts of data. Fortunately, simple methods to remove such an "under-sampling bias" have been known for a long time (Miller, 1955).

Of great interest in the 1960s was the measuring of the "capacity" of various senses. The procedure varied in detail, but was essentially the same: the subjects were asked to label stimuli (say, tones of different frequencies) with different numbers. The mutual information between the stimuli and the numbers assigned by the subjects was then calculated with different numbers of stimuli presented (see Figure 1.2). Given only two stimuli, a subject would almost never make a mistaken identification, but as the number of stimuli to be labelled increased, subjects started to make mistakes. By estimating where the function relating mutual information to the number of

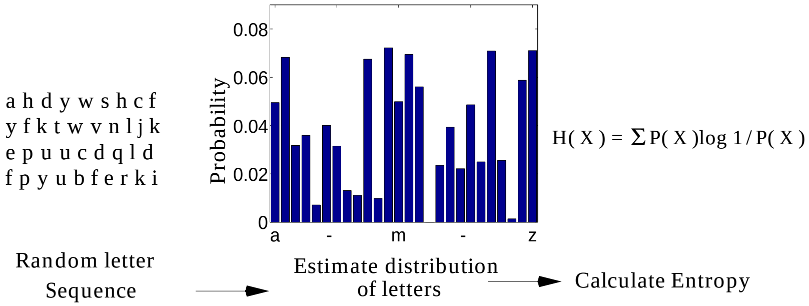

Figure 1.1. The most straightforward method to calculate entropy or mutual information is direct estimation of the probability distributions (after Baddeley, 1956). One case where this is appropriate is in using the entropy of subjects' random number generation ability as a measure of cognitive load. The subject is asked to generate random digit sequences in time with a metronome, either as the only task, or while simultaneously performing a task such as card sorting. Depending on the difficulty of the other task and the speed of generation, the "randomness" of the digits will decrease. The simplest way to estimate entropy is to estimate the probability of different letters. Using this measure of entropy, redundancy (entropy/maximum entropy) decreases linearly with generation time, and also with the difficulty of the other task. This has subsequently proved a very effective measure of cognitive load.

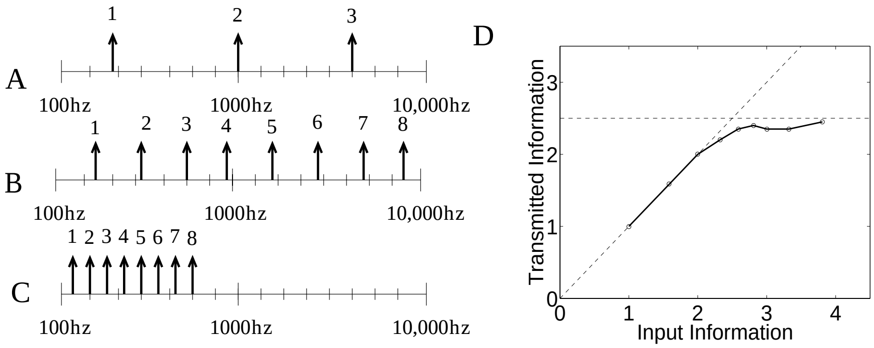

Figure 1.2. Estimating the "channel capacity" for tone discrimination (after Pollack, 1952, 1953). The subject is presented with a number of tones and asked to assign numeric labels to them. Given only three tones (A), the subject has almost perfect performance, but as the number of tones increase (B), performance rapidly deteriorates. This is not primarily an early sensory constraint, as performance is similar when the tones are tightly grouped (C). One way to analyse such data is to plot the transmitted information as a function of the number of input stimuli (D). As can be seen, up until about 2.5 bits, all the available information is transmitted, but when the input information is above 2.5 bits, the excess information is lost. This limited capacity has been found for many tasks and was of great interest in the 1960s.

input categories asymptotes, an estimate of subjects channel capacity can be made. Surprisingly this number is very small – about 2.5 bits. This capacity estimate approximately holds for a large number of other judgements: loudness (2.3 bits), tastes (1.9 bits), points on a line (3.25 bits), and this leads to one of the best titles in psychology – the "seven plus or minus two" of Miller (1956) refers to this small range (between 2.3 bits ($\log_2 5$) and 3.2 bits ($\log_2 9$)).

Again in these tasks, since the number of labels usable by subjects is small, it is very possible to directly estimate the probability distributions with reasonable amounts of data. If instead subjects were reliably able to label 256 stimuli (8 bits as opposed to 2.5 bits capacity), we would again get into problems of collecting amounts of data sufficient to specify the distributions, and methods based on the direct estimation of probability distributions would require vast amounts of subjects' time.

Continuous Distributions

Given that the data are discrete, and we have enough data, then simply estimating probability distributions presents few conceptual problems. Unfortunately if we have continuous variables such as membrane potentials, or reaction times, then we have a problem. While the entropy of a discrete probability distribution is finite, the entropy of any continuous variable is infinite. One easy way to see this is that using a single real number between 0 and 1, we could very simply code the entire *Encyclopedia Britannica*. The first two digits after the decimal place could represent the first letter; the second two digits could represent the second letter, and so on. Given no constraint on accuracy, this means that the entropy of a continuous variable is infinite.

Before giving up hope, it should be remembered that mutual information as specified by equation 1.4 is the *difference* between two entropies. It turns out that as long as there is some noise in the system ($H(X|Y) > 0$), then the difference between these two infinite entropies is finite. This makes the role of noise vital in any information theory measurement of continuous variables.

One particular case is if both the signal and noise are Gaussian (i.e. normally) distributed. In this case the mutual information between the signal (s) and the noise-corrupted version (s_n) is simply:

$$I(s; s_n) = \frac{1}{2}\log_2\left(1 + \frac{\sigma^2_{signal}}{\sigma^2_{noise}}\right) \tag{1.5}$$

where σ^2_{signal} is the variance of the signal, and σ^2_{noise} is the variance of the noise. This has the expected characteristics: the larger the signal relative to the noise, the larger the amount of information transmitted; a doubling of the signal will result in an approximately 1 bit increase in information transmission; and the information transmitted will be independent of the unit of measurement.

It is important to note that the above expression is only valid when both the signal and noise are Gaussian. While this is often a reasonable and testable assumption because of the central limit theorem (basically, the more things we add, usually the more Gaussian the system becomes), it is still only an estimate and can underestimate the information (if the signal is more Gaussian than the noise) or overestimate the information (if the noise is more Gaussian than the signal).

A second problem concerns correlated signals. Often a signal will have structure – for instance, it could vary only slowly over time. Alternatively, we could have multiple measurements. If all these measurements are independent, then the situation is simple – the entropies and mutual informations simply add. If, on the other hand, the variables are correlated across time, then some method is required to take these correlations into account. In an extreme case if all the measurements were identical in both signal and noise, the information from one such measurement would be the same as the combined information from all: it is important to in some way deal with these effects of correlation.

Perhaps the most common way to deal with this "correlated measurements" problem is to transform the signal to the Fourier domain. This method is used in a number of papers in this volume and the underlying logic is described in Figure 1.3.

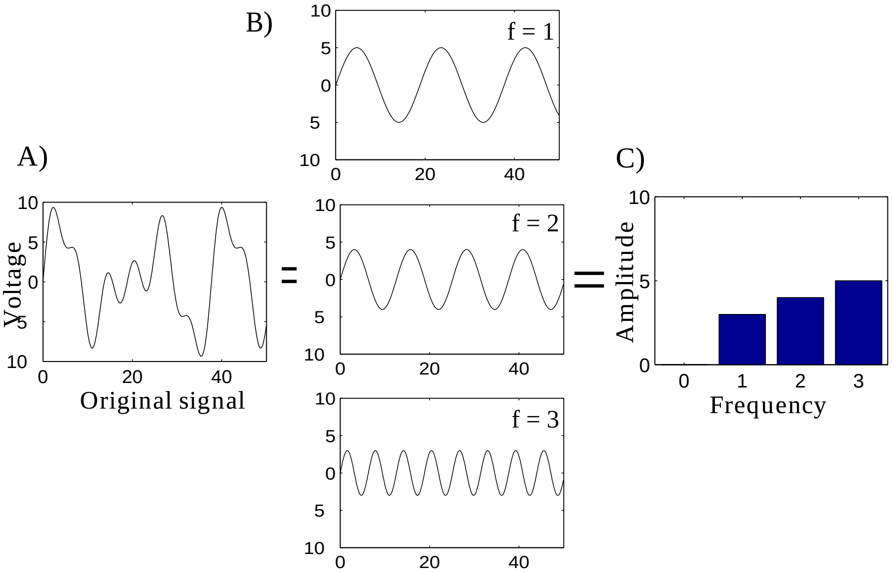

Figure 1.3. Taking into account correlations in data by transforming to a new representation. (A) shows a signal varying slowly as a function of time. Because the voltages at different time steps are correlated, it is not possible to treat each time step as independent and work out the information as the sum of the information values at different time steps. One way to approach this problem is to transform the signal to a new representation where all components are now uncorrelated. If the signal is Gaussian, transforming to a Fourier series representation has this property. Here we represent the original signal (A) as a sum of sines and cosines of different frequencies (B). While the individual time measurements are correlated, if the signal is Gaussian, the amounts of each Fourier components (C) will be uncorrelated. Therefore the mutual information for the whole signal will simply be the sum of the information values for the individual frequencies (and these can be calculated using equation 1.5).

The Fourier transform method always uses the same representation (in terms of sines and cosines) independent of the data. In some cases, especially when we do not have that much data, it may be more useful to choose a representation which still has the uncorrelated property of the Fourier components, but is optimised to represent a particular data set. One plausible candidate for such a method is principal components analysis. Here a new set of measurements, based on linear transformation of the original data, is used to describe the data. The first component is the linear combination of the original measurements that captures the maximum amount of variance. The second component is formed by a linear combination of the original measurements that captures as much of the variance as possible while being orthogonal to the first component (and hence independent of the first component if the signal is Gaussian). Further components can be constructed in a similar manner. The main advantage over a Fourier-based representation is that more of the signal can be described using fewer descriptors and thus less data is required to estimate the characteristics of the signal and noise. Methods based on principal-component-based representations of spikes trains have been applied to calculating the information transmitted by cortical neurons (Richmond and Optican, 1990).

All the above methods rely on an assumption of Gaussian nature of the signal, and if this is not true and there exist non-linear relationships between the inputs and outputs, methods based on Fourier analysis or principal components analysis can only give rather inaccurate estimates. One method that can be applied in this case is to use a non-linear compression method to generate a compressed representation before performing the information estimation (see Figure 1.4).

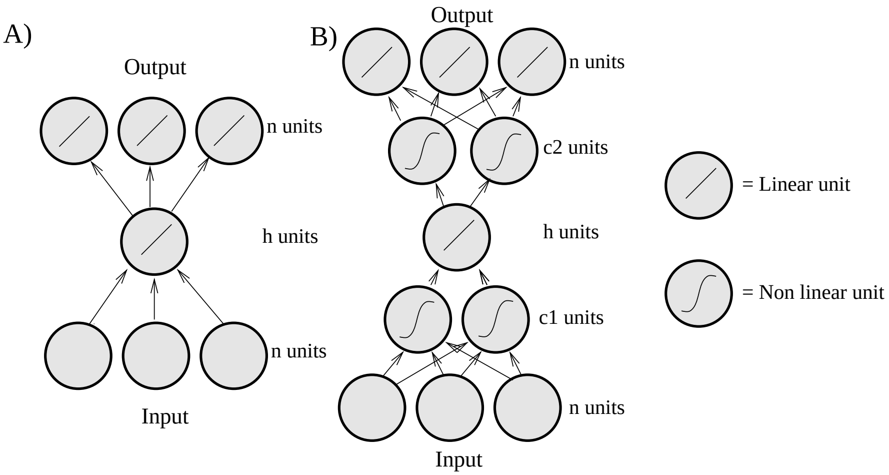

Figure 1.4. Using non-linear compression techniques for generating compact representations of data. Linear principal components analysis can be performed using the neural network shown in (A) where a copy of the input is used as the target output. On convergence, the weights from the *n* input units to the *h* coding units will span the same space as the first *h* principal components and, given that the input is Gaussian, the coding units will be a good representation of the signal. If, on the other hand, there is non-Gaussian non-linear structure in the signals, this approach may not be optimal. One possible approach to dealing with such non-linearity is to use a compression-based algorithm to create a non-linear compressed representation of the signals. This can be done using the non-linear generalisation of the simple network to allow non-linearities in processing (shown in (B)). Again the network is trained to recreate its input from its output, while transmitting the information through a bottleneck, but this time the data is allowed to be transformed using an arbitrary non-linearity before coding. If there are significant non-linearities in the data, the representation provided by the bottleneck units may provide a better representation of the input than a principal-components-based representation. (After Fotheringhame and Baddeley, 1997.)

Estimation Using an "Intelligent" Predictor

Though the direct measurement of the probability distributions is conceptually the simplest method, often the dimensionality of the problem renders this implausible. For instance, if interested in the entropy of English, one could get better and better approximations by estimating the probability distribution of letters, letter pairs, letter triplets, and so on. Even for letter triplets, there is the probability of $27^3 = 19{,}683$ possible three-letter combinations to estimate: the amount of data required to do this at all accurately is prohibitive. This is made worse because we know that many of the regularities of English would only be revealed over groups of more than three letters. One potential solution to this problem is available if we have access to a good model of the language or predictor. For English, one source of a predictor of English is a native speaker. Shannon (see Table 1.1) used this to devise an ingenious method for estimating the entropy of English as described in Table 1.1.

Even when we don't have access to such a good predictor as an English language speaker, it often simpler to construct (or train) a predictor rather than to estimate a large number of probabilities. This approach to estimating mutual information has been applied (Heller et al., 1995) to estimation of the visual information transmission properties of neurons in both the primary visual cortex (also called V1; area 17; or striate cortex) and the inferior temporal cortex (see Figure 1.5). Essentially the spikes generated by neurons when presented various stimuli were coded in a number of different ways (the

Table 1.1. *Estimating the entropy of English using an intelligent predictor (after Shannon, 1951).*

T	H	E	R	E		I	S		N	O		R	E	V	E	R	S	E
1	1	1	5	1	1	2	1	1	2	1	1	15	1	17	1	1	1	2
	O	N		A		M	O	T	O	R	C	Y	C	L	E			
1	3	2	1	2	2	7	1	1	1	1	4	1	1	1	1			

Above is a short passage of text. Underneath each letter is the number of guesses required by a person to guess that letter based only on knowledge of the previous letters. If the letters were completely random (maximum entropy and no redundancy), the best predictor would take on average 27/2 guesses (26 letters and a space) for every letter. If, on the other hand, there is complete predictability, then a predictor would only require only one guess per letter. English is between these two extremes and, using this method, Shannon estimated an entropy per letter of between 1.6 and 0.6 bits per letter. This contrasts with $\log 27 = 4.76$ bits if every letter was equally likely and independent. Technical details can be found in Shannon (1951) and Attneave (1959).

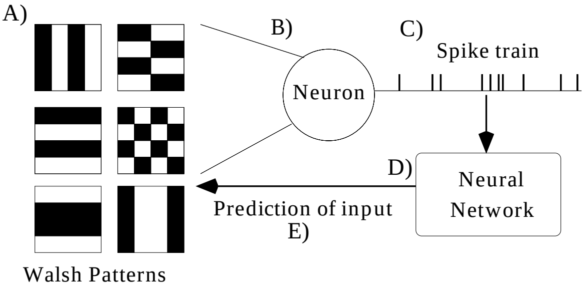

Figure 1.5. Estimating neuronal information transfer rate using a neural network based predictor (after Heller et al., 1995). A collection of 32 4×4 Walsh patterns (and their contrast reversed versions) (A) were presented to awake Rhesus Macaque monkeys, and the spike trains generated by neurons in V1 and IT recorded (B and C). Using differently coded versions of these spike trains as input, a neural network (D) was trained using the back-propagation algorithm to predict which Walsh pattern was presented. Intuitively, if the spike train contains a lot of information about the input, then an accurate prediction is possible, while if there is very little information then the spike train will not allow accurate prediction of the input. Notice that (1) the calculated information will be very dependent on the choice (and number of) of stimuli, and (2) even though we are using a predictor, implicitly we are still estimating probability distributions and hence we require large amounts of data to accurately estimate the information. Using this method, it was claimed that the neurons only transmitted small amounts of information ($\approx$ 0.5 bits), and that this information was contained not in the exact timing of the spikes, but in a local "rate".

average firing rate, vectors representing the presence and absence of spikes, various low-pass-filtered versions of the spike train, etc). These codified spike trains were used to train a neural network to predict the visual stimulus that was presented when the neurons generated these spikes. The accuracy of these predictions, given some assumptions, can again be used to estimate the mutual information between the visual input and the differently coded spike trains estimated. For these neurons and stimuli, the information transmission is relatively small ($\approx$ 0.5 bits s^{-1}).

Estimation Using Compression

One last method for estimating entropy is based on Shannon's coding theorem, which states that the smallest size that any compression algorithm can compress a sequence is equal to its entropy. Therefore, by invoking a number of compression algorithms on the sample sequence of interest, the smallest compressed representation can be taken as an upper bound on that sequence's entropy. Methods based on this intuition have been more common in genetics, where they have been used to ask such questions as does "coding" DNA have higher or lower entropy than "non-coding" DNA (Farach et al., 1995). (The requirements of quick convergence and reasonable computation

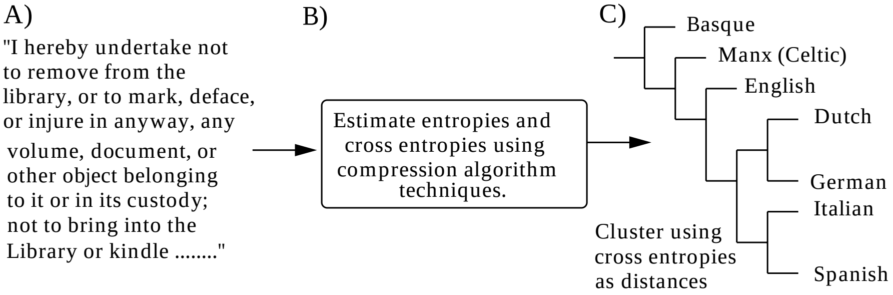
A)
"I hereby undertake not to remove from the library, or to mark, deface, or injure in anyway, any volume, document, or other object belonging to it or in its custody; not to bring into the Library or kindle ...."
B)
Estimate entropies and cross entropies using compression algorithm techniques.
C)
Basque
Manx (Celtic)
English
Dutch
German
Italian
Spanish
Cluster using cross entropies as distances

Figure 1.6. Estimating entropies and cross entropies using compression-based techniques. The declaration of the Bodleian Library (Oxford) has been translated into more than 50 languages (A). The entropy of these letter sequences can be estimated using the size of a compressed version of the statement. If the code book derived by the algorithm for one language is used to code another language, the size of the code book will reflect the cross entropy (B). Hierarchical minimum distance cluster analysis, using these cross entropies as a distances, can then be applied to this data (a small subset of the resulting tree is shown (C)). This method can produce an automatic taxonomy of languages, and has been shown to correspond very closely to those derived using more traditional linguistic analysis (Juola, P., personal communication).

time mean that only the earliest algorithms simply performed compression, but the concept behind later algorithms is essentially the same.)

More recently, this compression approach to entropy estimation has been applied to automatically calculating linguistic taxonomies (Figure 1.6). The entropy was calculated using a modified compression algorithm based on Farach et al. (1995). Cross entropy was estimated using the compressed length when the code book derived for one language was used to compress another. Though methods based on compression have not been commonly used in the theoretical neuroscience community (but see Redlich, 1993), they provide at least interesting possibilities.

1.5 Maximising Information Transmission

The previous section was concerned with simply measuring entropy and information. One other proposal that has received a lot of attention recently is the proposition that some cortical systems can be understood in terms of them maximising information transmission (Barlow, 1989). There are a number of reasons supporting such an information maximisation framework:

Maximising the Richness of a Representation. The richness and flexibility of the responses to a behaviourally relevant input will be limited by the number of different states that can be discriminated. As an extreme case, a protozoa that can only discriminate between bright and dark will have less flexible navigating behaviour than an insect (or human) that has an accurate representation of the grey-level structure of the visual world. Therefore, heuristically, evolution will favour representations that maximise information transmission, because these will maximise the number of discriminable states of the world.

As a Heuristic to Identify Underlying Causes in the Input. A second reason is that maximising information transmission is a reasonable principle for generating representations of the world. The pressure to compress the world often forces a new representation in terms of the actual "causes" of the images (Olshausen and Field, 1996a). A representation of the world in terms of edges (the result of a number of information maximisation algorithms when applied to natural images, see for instance Chapter 5), may well be easier to work with than a much larger and redundant representation in terms of the raw intensities across the image.

To Allow Economies to be Made in Space, Weight and Energy. By having a representation that is efficient at transmitting information, it may be possible to economise on some other of the system design. As described in Chapter 3, an insect eye that transmits information efficiently can be smaller and lighter, and can consume less energy (both when operating and when being transported). Such "energetic" arguments can also be applied to, say, the transmission of information from the eye to the brain, where an inefficient representation would require far more retinal ganglion cells, would take significantly more space in the brain, and use a significantly larger amount of energy.

As a Reasonable Formalism for Describing Models. The last reason is more pragmatic and empirical. The quantities required to work out how efficient a representation is, and the nature of a representation that maximises information transmission, are measurable and mathematically formalisable. When this is done, and the "optimal" representations compared to the physiological and psychophysical measurements, the correspondence between these optimal representations and those observed empirically is often very close. This means that even if the information maximisation approach is only heuristic, it is still useful in summarising data.

How then can one maximise information transmission? Most approaches can be understood in terms of a combination of three different strategies:

- Maximise *the number of effective measurements* by making sure that each measurement tells us about a different thing.
- Maximise the signal whilst *minimising the noise.*
- Subject to the external constraints placed on the system, *maximise the efficiency* of the questions asked.

Maximising the Effective Number of Questions

The simplest method of increasing information transmission is to increase the number of measurements made: someone asking 50 questions concerning the page flipped to in a book has more chance of identifying it than someone who asks one question. Again an eye connected by a large number of retinal ganglion cells to later areas should send more information than the single ganglion cell connected to an eyecup of a flatworm.

This insight is simple enough not to rely on information theory, but the raw number of measurements is not always equivalent to the "effective" number of measurements. If given two questions to identify a page in the book – if the first one was "Is it between pages 1 and 10?" then a second of "Is it between 2 and 11?" would provide remarkably little extra information. In particular, given no noise, the maximum amount of information can be transmitted if all measurements are independent of each other.

A similar case occurs in the transmission of information about light entering the eye. The outputs of two adjacent photoreceptors will often be measuring light coming from the same object and therefore send very correlated signals. Transmitting information to later stages simply as the output of photoreceptors would therefore be very inefficient, since we would be sending the same information multiple times. One simple proposal for transforming the raw retinal input before transmitting it to later stages is shown in Figure 1.7, and has proved successful in describing a number of facts about early visual processing (see Chapter 3).

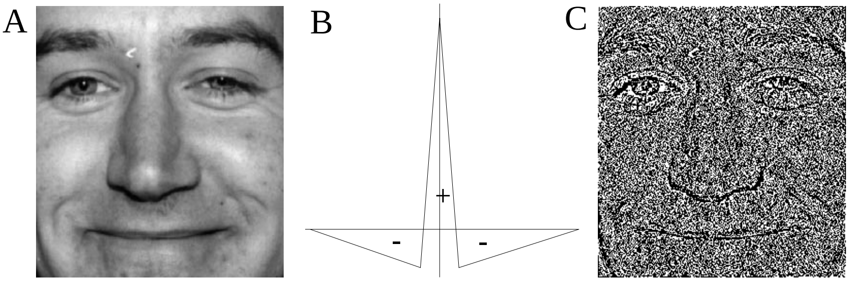

Figure 1.7. Maximising information transmission by minimising redundancy. In most images, (A) the intensity arriving at two locations close together in the visual field will often be very similar, since it will often originate from the same object. Sending information in this form is therefore very inefficient. One way to improve the efficiency of transmission is not to send the pixel intensities, but the difference between the intensity at a location and that predicted from the nearby photoreceptors. This can be achieved by using a centre surround receptive field as shown in (B). If we transmit this new representation (C), far less channel capacity is used to send the same amount of information. Such an approach seems to give a good account of the early spatial filtering properties of insect (Srinivasan et al., 1982; van Hateren, 1992b) and human (Atick, 1992b; van Hateren, 1993) visual systems.

Guarding Against Noise

The above "independent measurement" argument is only true to a point. Given that the person you ask the question of speaks clearly, then ensuring that each measurement tells you about a different thing is a reasonable strategy. Unfortunately, if the person mumbles, has a very strong accent, or has possibly been drinking too much, we could potentially miss the answer to our questions. If this happens, then because each question is unrelated to the others, an incorrect answer cannot be detected by its relationship to other questions, nor can they be used to correct the mistake. Therefore, in the presence of noise, some redundancy can be helpful to (1) detect corrupted information, and (2) help correct any errors. As an example, many non-native English speakers have great difficulty in hearing the difference between the numbers 17 and 70. In such a case it actually might be worth asking "is the page above seventy" as well as "is it above fifty" since this would provide some guard against confusion of the word seventy. This may also explain the charming English habit of shouting loudly and slowly to foreigners.

The appropriate amount of redundancy will depend on the amount of noise: the amount of redundancy should be high when there is a lot of noise, and low when there is little. Unfortunately this can be difficult to handle when the amount of noise is different at different times, as in the retina. Under a bright illuminant, the variations in image intensity (the signal) will be much larger than the variations due to the random nature of photon arrival or the unreliability of synapses (the noise). On the other hand, for very low light conditions this is no longer the case, with the variations due to the noise now relatively large. If the system was to operate optimally, the amount of redundancy in the representation should change at different illumination levels. In the primate visual system, the spatial frequency filtering properties of the "retinal filters" change as a function of light level, consistent with the retina maximising information transmission at different light levels (Atick, 1992b).

Making Efficient Measurements

The last way to maximise information transmission is to ensure not only that all measurements measure different things, and noise is dealt with effectively, but also that the measurements made are as informative as possible, subject to the constraints imposed by the physics of the system.

For binary yes/no questions, this is relatively straightforward. Consider again the problem of guessing a page in the *Encyclopedia Britannica*. Asking the question "Is it page number 1?" is generally not a good idea – if you happen to guess correctly then this will provide a great deal of information (technically known as suprisal), but for the majority of the time you will know very little more. The entropy (and hence the maximum amount of information transmission) is maximal when the uncertainty is maximal, and this occurs when both alternatives are equally likely. In this case we want questions where "yes" is has the same probability as "no". For instance a question such as "Is it in the first or second half of the book?" will generally tell you more than "Is it page 2?". The entropy as a function of probability is shown for a yes/no system (binary channel) in Figure 1.8.

When there are more possible signalling states than true and false, the constraints become much more important. Figure 1.9 shows three of the simplest cases of constraints and the nature of the outputs (if we have no noise) that will maximise information transmission. It is interesting to note that the spike trains of neurons are exponentially distributed as shown in Figure 1.9(C), consistent with maximal information transmission subject to an average firing rate constraint (Baddeley et al., 1997).

1.6 Potential Problems and Pitfalls

The last sections were essentially positive. Unfortunately not all things about information theory are good:

The Huge Data Requirement. Possibly the greatest problem with information theory is its requirement for vast amounts of data if the results are to tell us more about the data than about the assumptions used to calculate its value. As mentioned in Section 1.4, estimating the probability of every three-letter combination in English would require sufficient data to estimate 19,683 different probabilities. While this may actually be possible given the large number of books available electronically, to get a better approximation to English, (say, eight-letter combinations), the amount of data required

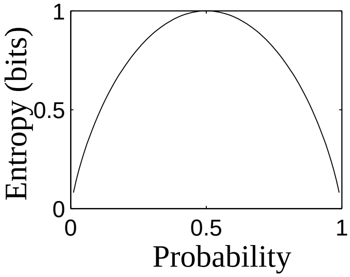

Figure 1.8. The entropy of a binary random (Bernoulli) variable is a function of its probability and maximum when its probability is 0.5 (when it has an entropy of 1 bit). Intuitively, if a measurement is always false (or always true) then we are not uncertain of its value. If instead it is true as often as not, then the uncertainty, and hence the entropy, is maximised.

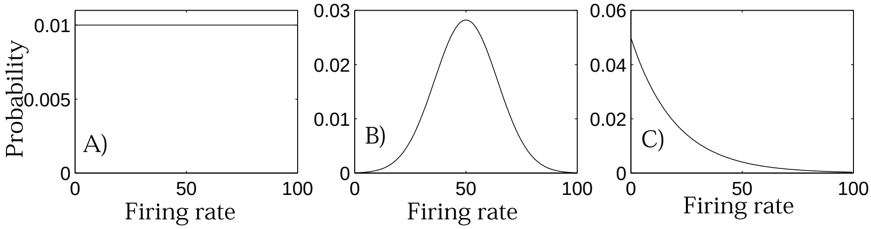

Figure 1.9. The distribution of neuronal outputs consistent with optimal information transmission will be determined by the most important constraints operating on that neuron. First, if a neuron is only constrained by its maximum and minimum output, then the maximum entropy, and therefore the maximum information that could be transmitted, will occur when all output states are equally likely (A) (Laughlin, 1981). Second, a constraint favoured for mathematical convenience is that the power (or variance) of the output states is constrained. Given this, entropy is maximised for a Gaussian firing rate distribution (B). Third, if the constraint is on the average firing rate of a neuron, higher firing rates will be more "costly" than low firing rates, and an exponential distribution of firing rates would maximise entropy (C). Measurements from V1 and IT cells show that neurons in these areas have exponentially distributed outputs when presented with natural images (Baddeley et al., 1997), and hence are at least consistent with maximising information transmission subject to an average rate constraint.

becomes completely unrealistic. Problems of this form are almost always present when applying information theory, and often the only way to proceed is to make assumptions which are possibly unfounded and often difficult to test. Assuming true independence (very difficult to verify even with large data sets), and assuming a Gaussian signal and noise can greatly cut down on the number of measurements required. However, these assumptions often remain only assumptions, and any interpretations of the data rest strongly on them.

Information and Useful Information. Information theory again only measures whether there are variations in the world that can be reliably discriminated. It does not tell us if this distinction is of any interest to the animal. As an example, most information-maximisation-based models of low-level vision assume that the informativeness of visual information is simply based on how much it varies. Even at the simplest level, this is difficult to maintain as variation due to, say, changes in illumination is often of less interest than variations due to changes in reflectance, while the variance due to changes in illumination is almost always greater than that caused by changes in reflectance. While the simple "variation equals information" may be a useful starting point, after the mathematics starts it is potentially easy to forget that it is only a first approximation, and one can be led astray.

Coding and Decoding. A related problem is that information theory tells us if the information is present, but does not describe whether, given the computational properties of real neurons, it would be simple for neurons to extract. Caution should therefore be expressed when saying that information present in a signal is information available to later neurons.

Does the Receiver Know About the Input? Information theory makes some strong assumptions about the system. In particular it assumes that the receiver knows everything about the statistics of the input, and that these statistics do not change over time (that the system is stationary). This assumption of stationarity is often particularly unrealistic.

1.7 Conclusion

In this chapter it was hoped to convey an intuitive feel for the core concepts of information theory: entropy and information. These concepts themselves are straightforward, and a number of ways of applying them to calculate information transmission in real systems were described. Such examples are intended to guide the reader towards the domains that in the past have proved amenable to information theoretic techniques. In particular it is argued that some aspects of cortical computation can be understood in the context of maximisation of transmitted information. The following chapters contain a large number of further examples and, in combination with Cover and Thomas (1991) and Rieke et al. (1997), it is hoped that the reader will find this book helpful as a starting point in exploring how information theory can be applied to new problem domains.