Network: Comput. Neural Syst. **12** (2001) 441–472 PII: S0954-898X(01)25412-8

Neural coding and decoding: communication channels and quantization

Alexander G Dimitrov and John P Miller

Center for Computational Biology, Montana State University, Bozeman, MT 59717-3148, USA

E-mail: alex@cns.montana.edu and jpm@cns.montana.edu

Received 19 July 2000, in final form 15 March 2001
Published 15 August 2001
Online at stacks.iop.org/Network/12/441

Abstract
We present a novel analytical approach for studying neural encoding. As a first step we model a neural sensory system as a communication channel. Using the method of typical sequence in this context, we show that a coding scheme is an almost bijective relation between equivalence classes of stimulus/response pairs. The analysis allows a quantitative determination of the type of information encoded in neural activity patterns and, at the same time, identification of the code with which that information is represented. Due to the high dimensionality of the sets involved, such a relation is extremely difficult to quantify. To circumvent this problem, and to use whatever limited data set is available most efficiently, we use another technique from information theory—quantization. We quantize the neural responses to a reproduction set of small finite size. Among many possible quantizations, we choose one which preserves as much of the informativeness of the original stimulus/response relation as possible, through the use of an information-based distortion function. This method allows us to study coarse but highly informative approximations of a coding scheme model, and then to refine them automatically when more data become available.

1. Introduction

One of the steps toward understanding the neural basis of an animal's behaviour is characterizing the code with which its nervous system represents information. All computations underlying an animal's behavioural decisions are carried out within the context of this code. A determination of the neural coding schemes is an extremely important goal, due not only to our interest in the nature of the code itself, but also to the constraints that this knowledge places on the development of theories for the biophysical mechanisms underlying neural computation [19].

0954-898X/01/040441+32$30.00 © 2001 IOP Publishing Ltd Printed in the UK

Deciphering the neural code of a sensory system is often reduced to a few interconnected tasks. One task is to determine the correspondence between neural activity and sensory signals. That task can be reduced further to two interrelated problems: determining the specific stimulus parameters or features encoded in the neural ensemble activity and determining the nature of the neural symbols with which that information is encoded. An ancillary task is to define quantitative measures of significance with which the sensory information and associated neural symbols are correlated. Considerable progress has been made by approaching these tasks as independent problems. Approaches that we and others have taken include stimulus reconstruction [26, 35] and the use of impoverished stimulus sets to characterize stimulus/response properties [15,36]. Recent work also provided direct estimates of information-theoretic measures of correlations between stimulus and response [5,33], while completely bypassing the problem of identifying the stimulus. However, independent treatment of these interconnected tasks often introduces multiple assumptions that prevent their complete solution. Some of these approaches start with an assumption of the relevant codewords (e.g. a single spike in the first-order stimulus reconstruction method, or the mean spike rate over a defined interval) and proceed by calculating the expected stimulus features that are correlated with these codewords. Other approaches make an assumption about the relevant stimulus features (e.g. moving bars and gratings in the investigation of parts of the visual cortex) and proceed to identify the pattern of spikes that follow the presentation of these features. We have developed an analytical approach that takes all three tasks into consideration simultaneously. Specifically, the aim of this approach is to allow a quantitative determination of the type of information encoded in neural activity patterns and, at the same time, identify the code with which that information is represented.

There are two specific goals of this paper. The first is to formulate a model of a neural sensory system as a communication channel (section 2). In this context we show that a coding scheme consists of classes of stimulus/response pairs which form a structure akin to a dictionary: each class consists of a stimulus set and a response set, which are synonymous. The classes themselves are almost independent, with few intersecting members. The second goal is to design a method for discovering this dictionary-like structure (section 3). To do this, we quantize the neural responses to a reproduction set of small finite size. Among many possible quantizations, we choose one which preserves as much of the informativeness of the original stimulus/response relation as possible, through the use of an information-based distortion function.

We start with the observation that the neural code must satisfy two conflicting demands. On one hand, the organism must recognize the same natural object as identical in repeated exposures. In this sense, the signal processing operations of the organism need to be deterministic at this 'behavioural' level. On the other hand, the neural coding scheme must deal with projections of the sensory environment to a smaller stimulus space and uncertainties introduced by external and internal noise sources. Therefore, the neural signal processing must, by necessity, be stochastic on a finer scale. In this light, the functional issues that confront the early stages of any biological sensory system are very similar to the issues encountered by communication engineers in their work of transmitting messages across noisy media.

With this in mind we model the input/output relationship present in a biological sensory system as a communication channel [31]. Although this approach has been suggested before [2, 3], to our knowledge all the properties that information theory assigns to this object have not been completely appreciated in the neural research literature. The principal method that we present is based on the identification of jointly typical sequences ([6, appendix A.4]) in the stimulus/response sets. Joint typicality is a rigorously defined concept used extensively in information theory [6] and described in more detail in section 2.3. We use this technique to elucidate the structure of a neural stimulus/response relation. Although our coding model is stochastic, we demonstrate how an almost deterministic relation emerges naturally on the level of clusters of stimulus/response pairs.

To investigate such a model we consider the possibility that temporal patterns of spikes across a small ensemble of cells are the basic elements of information transmission in such system. Due to the high dimensionality of the sets involved, such a relation is extremely difficult to quantify [17, 33, 38]. To circumvent this problem, and to use whatever limited data set is available most efficiently, we quantize the neural responses to a reproduction set of small finite size (section 3.1). Quantization is another standard technique from information theory ([6, 14], appendix A.5.2). By quantizing to a reproduction space, the size of which is sufficiently small, we can assure that the data size requirements are much diminished.

Among many possible quantizations, we choose one which preserves as much of the informativeness of the original stimulus/response relation as possible, through the use of an information-based distortion function (section 3.2). The relationship between stimulus and reproduction will be an approximation of the coding scheme described above. This method allows us to study coarse but highly informative models of a coding scheme, and then to automatically refine them when more data become available. Ultimately, a simple description of the stimulus/response relation can be recovered (section 3.4).

2. Neural information processing

2.1. Neural systems as communication channels

Communication channels characterize a relation between two random variables: an input and an output. For neural systems, the output space is usually the set of activities of a group of neurons. The input space can be sensory stimuli from the environment or the set of activities of another group of neurons. We would like to recover the correspondence between stimuli and responses, which we call a *coding scheme* [34]. For any problem that has inputs and outputs, a coding scheme is a map (correspondence) from the input space to the output space. A decoding scheme is the inverse map. In general, both maps can be probabilistic and many to one. We will provide a more precise definition, which describes the situation when they can be considered almost deterministic and almost bijective.

The early stages of neural sensory processing encode information about sensory stimuli into a representation that is common to the whole nervous system. We will consider this encoding process within a probabilistic framework [3, 8, 18, 26]: *the signal X* is produced by a source with a probability $p(x)$. For example this may be a sensory stimulus from the animal's environment, or the activity of a set of neurons. *The encoder $q(y|x)$* is a stochastic map from one stochastic signal to another. This will model the operations of a neuronal layer. *The output signal Y* is produced by q with probability $p(y)$. This is the temporal pattern of activity across a set of cells. This description is naturally cast in the formalism of information theory. If we consider a neuron or a group of neurons as a communication channel, we can apply the theory almost directly for insights into the operation of a neural sensory system, including a detailed picture of the correspondence between sensory stimuli and their neural representations.

Our basic assumption is that sensory stimuli are represented by patterns of spikes in one or several neurons. The number of neurons comprising an information channel, the temporal extent of the patterns and temporal precision of spikes are parameters which can be determined from data [9]. We can formulate hypotheses about particular coding schemes by fixing the values of these parameters. Our initial assumption is that all distinct patterns of fixed length and precision may be important. Some ideas about information transmission in the presence of noise allow us to group patterns into larger classes and consider coding with these equivalence classes. This scheme is general, and encompasses most of the existing hypotheses about the nature of neural coding as special cases. Once equivalence classes are uncovered, they can be further analysed with regard to formal rules and physiological mechanisms which can describe them more succinctly. In certain cases these may turn out to be one of the above commonly considered coding schemes.

Sensory stimuli and their neural representation can be quite complex. Information theory suggests ways for dealing with this complexity by extracting the essential parts of the signal while maintaining most of its information content. To achieve this, we use the method of *typical sequences*. What follows is an informal discussion of their properties in relation to coding and our model of a sensory system. Many terms here are preceded by 'almost', 'nearly', 'close to' etc, which describe events that occur with high probability. The following three sections (2.2–2.4) discuss standard topics from information theory. See appendix A.3 for formal definitions and properties and [6] for a more complete treatment.

2.2. *Typical sequences*

The basic concepts of information theory are the entropy $H(X)$ and the mutual information $I(X, Y)$ of random sources (X, Y) ([6, 31], appendix A.2). When used in the analysis of communication systems, they have very specific meaning [31]. Consider a random source X. All the sequences of length n form the nth *extension* X^n of X. X^n can be described well by using about $2^{nH(X)}$ distinct messages. These are the *typical sequences* of X^n. Typical sequences (events) comprise the typical set and are nearly equiprobable. The typical set of X has probability near unity, and the number of elements in it is nearly $2^{nH(X)}$ (theorem appendix A.2). We can summarize this by saying 'Almost all events are almost equally surprising' [6]. This enables us to divide the set of all sequences into two sets—the *typical set*, where the entropy of a sequence is close to the entropy of the whole set, and the nontypical set, which contains the rest of the sequences. Any property that is proved for the typical sequences will be true with high probability and will determine the average behaviour of a large sample of sequences. The functional property of the entropy $H(X)$ here is to give the approximate size of the 'relevant' set of events in X^n.

To illustrate, consider $x \in X \equiv \{0, 1\}$, $p(x) = [p, q]$, $p + q = 1$ (the Bernoulli source). An example of a particular sequence of length 12 is (010001011010). The *n-typical sequences* are all the sequences of length n with approximately np zeros and nq ones. Notice that, despite its simplicity, this process can be used as a model of neural activity, where the presence of a spike is marked by one and its absence by zero. Spikes in a sequence are usually not independent of each other [22], so this should be considered at most a zeroth-order approximation.

Typical sequences are applicable to continuous random variables as well. Let $x \in \mathcal{N}(0, \sigma)$ (x is drawn from a normally distributed random variable with mean zero and variance σ^2). An n-sequence of these variables $(x_1, \ldots, x_n) \equiv X^n \in N(0^n, \sigma I^n)$ is obviously drawn from an n-dimensional normal distribution. The n-typical sequences here are all $X^n : \|X^n\|_2 \leqslant \sqrt{n(\sigma^2 + \epsilon)}$, that is all the points contained in an n-ball with the given radius. This property can be extended to general normal distributions, with a full covariance matrix, in which case the typical set lies within the n-ellipsoid determined by the principal directions and eigenvalues of the covariance matrix.

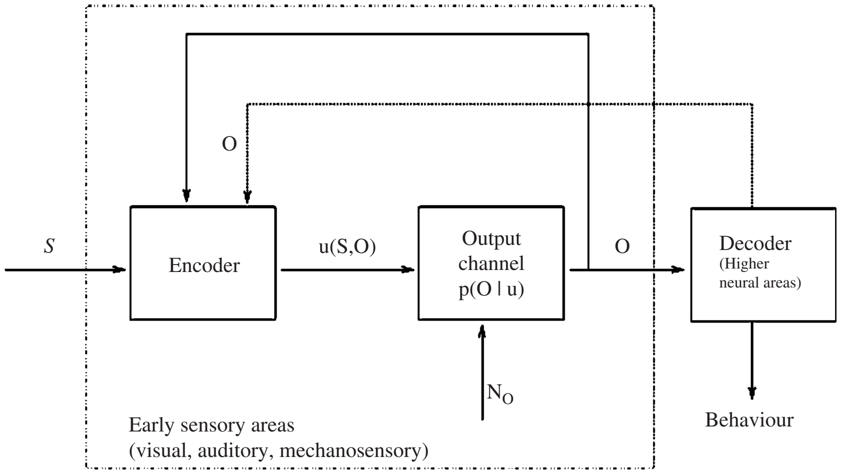
S
Encoder
u(S,O)
Output channel p(O | u)
O
O
Decoder (Higher neural areas)
N_O
Behaviour
Early sensory areas (visual, auditory, mechanosensory)

Figure 1. Neural sensory systems as information channels.

2.3. *Jointly typical sequences*

When analysing information channels, we deal with two sets of random sequences—input and output. In this case it is necessary to consider the combined behaviour of the pair (X, Y). We approach this by using *jointly typical sequences* ([6], appendix A.4). For a pair of sources (X, Y), there are about $2^{nH(X,Y)}$ *jointly typical* sequences of pairs (typical in the product space (X^n, Y^n)). As with typical sequences, all the elements are nearly equiprobable and the set has probability close to unity (appendix A.4). The correspondence between input and output sequences is not necessarily one to one, due to noise or mismatch between the stimulus and response sizes.

Not all pairs of typical x^n and typical y^n are also jointly typical. For each typical sequence in X^n, there are about $2^{nH(Y|X)}$ sequences in Y^n which are jointly typical with it, and vice versa. The probability that a randomly chosen pair is jointly typical is about $2^{-nI(X;Y)}$ [6]. Hence, for a fixed y^n, we can consider about $2^{nI(X;Y)}$ x^n before we are likely to come across a jointly typical pair. This suggests there are about $2^{nI(X;Y)}$ distinguishable messages (codewords) in X^n that can be communicated through Y^n (figure 2). Thus the knowledge of $I(X; Y)$ places important bounds on the performance of any coding scheme.

The structure of a coding scheme in this framework can be seen intuitively by the following argument [6].

> For each (typical) input n-sequence x^n, there are about $2^{nH(Y|X)}$ possible Y (jointly typical) sequences, all of which are approximately equally likely. The total number of possible (typical) Y sequences is about $2^{nH(Y)}$. In order to insure that no two X sequences produce the same Y output sequence, we need to divide the output set into chunks of size about $2^{nH(Y|X)}$, corresponding to the different input X sequences. The total number of disjoint sets after this procedure is about $2^{n(H(Y)-H(Y|X))} = 2^{nI(X;Y)}$. Hence we can transmit about $2^{nI(X;Y)}$ distinguishable X sequences of length n.

2.4. *Coding and decoding with jointly typical sequences*

Let the jointly typical pairs (x^n, y^n) represent related stimulus/response signals. Since there are $2^{nI(X;Y)}$ distinguishable codewords and $2^{nH(X,Y)}$ signals, some of the signals represent the same

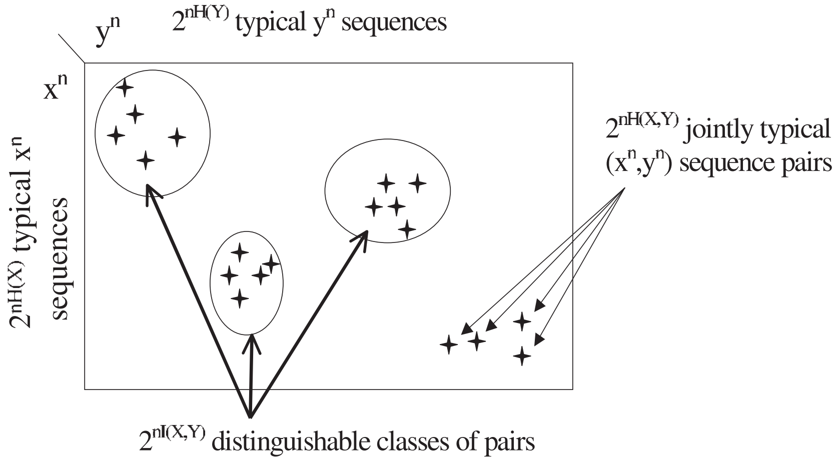
y^n
$2^{nH(Y)}$ typical y^n sequences
x^n
$2^{nH(X)}$ typical x^n sequences
$2^{nH(X,Y)}$ jointly typical (x^n,y^n) sequence pairs
$2^{nI(X,Y)}$ distinguishable classes of pairs

Figure 2. The structure of the jointly typical set. There are about $2^{nH(X)}$ typical stimulus (x) sequences and $2^{nH(Y)}$ typical response (y) sequences but only $2^{nH(X,Y)}$ jointly typical sequences. This suggests there are about $2^{nI(X;Y)}$ distinguishable *equivalence classes* C_i of stimulus/response pairs. Each class has about $2^{nH(X|Y)}$ input sequences and $2^{nH(Y|X)}$ output sequences. The number of stimulus/response pairs in each class is about $|C_i| \approx 2^{n(H(Y|X)+H(X|Y))}$.

codeword. Signals representing the same codewords can be combined in *equivalence classes*, which we call *codeword classes*. The codeword classes will represent the distinguishable messages that can be transmitted in this communication system. Within each class, a stimulus in X^n invokes a corresponding jointly typical response in Y^n with high probability (about $1 - 2^{-nI(X;Y)}$).

We define the *codebook* of this system as the map $\mathcal{F} : x^n \to y^n$. The codebook is stochastic on individual elements, so it is represented through the association probabilities $q(y^n|x^n)$. However, when considered on codeword classes, the map is *almost bijective*. That is, with probability close to unity, elements of Y^n from one codeword class are associated with elements of X^n in the same codeword class. We shall decode an output y^n as (any of) the inputs that belong to the same codeword class. Similarly, we shall consider the representation of an input x^n to be any of the outputs in the same codeword class. Stimuli from the same equivalence class are considered indistinguishable from each other, as are responses from the same class. If the stimulus or response has additional structure (e.g. resides in a metric space with a natural distance), the classes can be represented more succinctly using constraints from within this structure (e.g. by an exemplar with a minimum reconstruction error according to this distance).

An example of coding with jointly typical sequences. We shall use the binary source $X : x \in \{0, 1\}$, $p(x) = [r, 1 - r]$ and sequences from its finite extensions X^n. Here $H(p) = -r \log r - (1-r) \log(1-r)$. The communication system we consider in this example is the binary symmetric channel (BSC) (figure 3(*a*)). BSC is a model of a communication system with noise. The input and output are binary digits (zero or one; $x \in \mathcal{Z}_2$). The channel is described by the transition probabilities $\Pr(a \to a) = p$ for correct transmittal and $\Pr(a \to b) = q$ for an error (flipped bit); $p + q = 1, a, b \in [0, 1]$.

The BSC has a nice representation in terms of typical sequences. It can be considered as a random binary source with probability measure $[p, q]$ which generates sequences of ones and zeros. These sequences are then summed modulo 2 (xor-ed) with the original source to

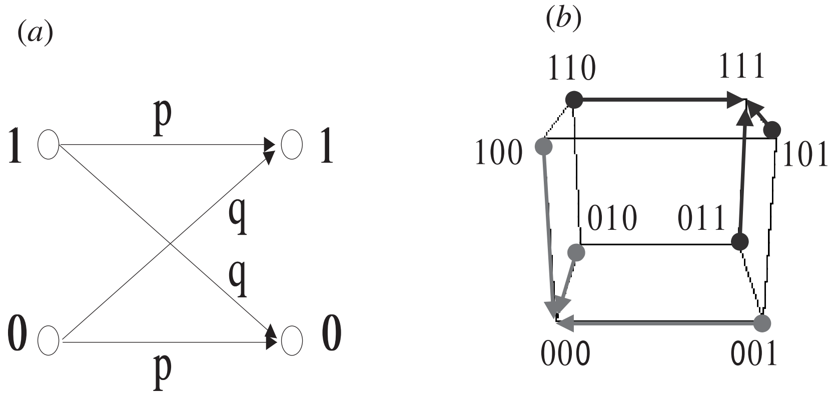

Figure 3. (*a*) The BSC. (*b*) A three-dimensional binary space X^3 and a particular coding scheme that allows the correction of one flipped bit. In this example, the codewords are 000 and 111 and the arrows mark the decoding process. Any pair of opposing vertices provides the same functionality.

produce the output. From the theorems on typicality (appendix A.3), the channel will typically produce sequences of nq ones (errors) and np zeros (correct transmission). There are about $2^{nH(p)}$ possible errors per transmitted input sequence; i.e., $H(Y^n|X^n) \approx nH(p)$.

A natural error measure for the BSC is the Hamming distance, which returns the number of bits in which two sequences differ. Figure 3(*b*) illustrates an example in X^3 which can correct one error in a block of three. The two distinct sets of arrows denote the two binary balls of unit radius in which this code can correct an error. We can use the two opposing vertices as codewords to transmit one bit of information. There are two equivalence classes with four elements each. Each element is decoded as the centre of the binary ball, with Hamming distance zero from the original codeword.

For X^n, let us consider only the uniform input distribution $p(x) = [1/2, 1/2]$. Then all sequences are typical. There are $2^{nH(X)} = 2^n$ of them. BSC with error probability q will produce about $2^{nH(q)}$ typical sequences of about nq flips. Thus each possible sequence in X will be received as one of about $2^{nH(q)}$ possible neighbours in Y. In order to transmit with small probability of error we need to pick as codewords sequences in X which are at least $nH(q)$ bits apart in Hamming distance. There are about $2^n/2^{nH(q)} = 2^{n(1-H(q))}$ such sequences.

2.5. Comments on the continuous case

The picture that we present is essentially one of discrete finite sets and relations between them. Some readers may raise the issue of how this whole scheme would work when both input and output processes are continuous, in space and/or time. We argue that most such cases can be handled by the current scheme. First, notice that any continuous coding scheme will contain uncertainties due to channel noise and measurement error at the receptor level. The scheme can be approximated to an arbitrary level of detail by a discrete coding scheme through the process of quantization (appendix A.5.2), which comes with another set of uncertainties (quantization noise). The two schemes will be indistinguishable in practice if their functionalities lie within each others' error bars. If they are distinguishable, there will be an experiment that can resolve the issue.

Even if we insist on considering continuous stimuli and responses, there are some arguments which again point to the benefit of discrete coding schemes. In the case of object and feature recognition, it was noted [24] that the continuity of stimuli is usually due to symmetries in the signal, which do not contribute to (and even interfere with) the recognition of features. The solution suggested in [24] is to pre-process the stimulus to remove as much of the continuous symmetries as possible and then to continue with the recognition algorithm.

Another argument comes from recent work in rate distortion theory. It was shown [27] that the optimal reproduction space of a continuous source is continuous only in a few special cases (e.g. if the source is Gaussian). In any other case, compressing and transmitting a continuous source is done best through discrete representatives.

We do not argue that this analytical approach covers every conceivable case. Our intention is to illustrate how coding with discrete sets can be used in a much wider context than may initially be perceived.

2.6. Neural coding with jointly typical sequences

The picture of coding and decoding with jointly typical sequences gives us a framework within which to think about the problem of neural coding and decoding. The description above (figure 2) is an idealization used to prove properties of optimal communication coding schemes, but the basic structure is valid for any coding scheme.

- A bijective or almost bijective coding scheme is needed for robust decoding, otherwise the animals will be confusing distinct stimuli.
- In any real system, there is noise which compromises the amount of information that can be represented. To achieve robustness, some of the capabilities of the system must be devoted to combating noise.
- Noise reduction can be achieved by combining signals that are likely to be confused into equivalence classes.
- A coding scheme that is almost bijective on equivalence classes will present a consistent representation of a sensory modality.

Unlike the idealized case of coding with jointly typical sequences, the codeword classes in a neural system need not be similar in size. In our model they still have the property of being almost self-consistent. Stimuli belonging to a certain codeword class will invoke a response within the same codeword class with high probability and only rarely produce a response outside this codeword class (in another class or nontypical). In such a case the stimulus will be considered incorrectly represented.

It should be noted that in the idealized picture there are many coding schemes which are optimal in the sense that the error rate asymptotically approaches zero [6]. This can be achieved by splitting the input and output into clusters of size appropriate for the channel ($H(X|Y)$ for the input, $H(Y|X)$ for the output), and then connecting them at random. When applied to the neural case, this means that a neural coding scheme for similar sensory modalities could be unique for species or even individuals.

Many neural coding schemes currently in use can be seen as special cases of the general coding scheme we describe here. For example, in rate coding schemes [1], all sequences in a short interval that have the same number of spikes are considered to be in the same equivalence class. The codeword classes are labelled by the number of spikes. All stimuli that precede a sequence in the same codeword class are considered jointly typical with it and hence decoded as being represented by this class. There is no guarantee that the classes are nonoverlapping or that the decoding error is small. Similarly, in population vector coding schemes [12, 29] the output sequences are assigned to classes based on the value of a linear functional of the number of spikes in each neuron (population mean). Decoding is done as before and labelled by the class identity, which is interpreted in [12] as the direction of planned hand movement. A spike latency code is an example of a continuous coding scheme. It fits in this formalism through the process of quantization (section 2.5). Classes are determined by the mean latency and jitter (variance) of the spike timing. The classes may be partially overlapping. A stimulus feature is decoded as in the rate code case, based on the latency class into which a spike falls.

Given this model of a neural subsystem, our task is to recover the codebook. There are two basic ways to approach this. Since the codebook is almost deterministic on equivalence classes, we can approximate it with a deterministic map. Then each equivalence class will have identified fixed members. All errors will be caused by disregarding certain regions of the event space. A more general approach is to represent $\mathcal{F}$ by the conditional probability of class membership $q(y|x)$ in which case we explicitly model the probability of misclassification, but lose some of the simplicity offered by jointly typical sequences. We shall use both approaches as warranted.

3. Finding the codebook

With communication systems, the usual use of information theory is to design a coding scheme given the structure of the communication channel. Our application differs, since now we are analysing a neural coding scheme already implemented within an animal's nervous system. Our goal is to uncover the structure described so far (section 2.4, figure 2) based on observations of the stimulus and response properties of the neural system under investigation.

3.1. Quantizing the response

As we mentioned before, the information quantities H and I depend only on the underlying probability function and not on the structure of the event space. This allows us to estimate them in cases where more traditional statistical measures (e.g. variance, correlations etc) simply do not exist. There is a drawback, though, since now we have to model the necessary probabilities, or else use large numbers of data to estimate the probabilities nonparametrically. The problem is compounded by several factors. Recent research suggests that nervous systems represent sensory stimuli by using relatively long temporal windows (tens to a few hundred ms in diverse preparations) [9, 17, 23, 25] and through the coordinated activity of multiple neurons [17, 21, 30, 39]. Unlike research which is geared towards finding better unbiased estimates of H and I [33, 38], our goal is to recover the complete coding scheme used in a sensory system. As pointed out in [17], the number of data points needed for nonparametric analysis of neural responses which are correlated across long time periods (length T) and multiple neurons (K) grows exponentially with T and K. It is conceivable that for some systems the required data recording time may well exceed the expected lifespan of the system.

To resolve this issue we choose to sacrifice some detail in the description of the coding scheme in order to obtain robust estimates of a coarser description. This can be achieved through quantization (appendix A.5.2, [6, 14]) of the neural representation Y into a coarser representation in a smaller event space Y_N. Y_N is referred to as the *reproduction* of Y. By controlling the size of the reproduction, we ensure that the data requirements to describe such a relation are much diminished. Instead of exponentially growing with T and K, now the number of needed data is proportional to N, the size of the reproduction, which is chosen by the researcher to be small. Most of the results in this section are valid for the general case when the sources are continuous, ergodic random variables [14]. However, the formulation for the most general case requires special attention to details. For clarity of the presentation, we shall assume that all random variables are finite (though possibly large) and discrete.

Quantizers are maps from one probability space to another. They can be deterministic (functions) or stochastic (given through a conditional probability) (appendix A.5.2). The size of the reproduction space is smaller than the size of the quantized space. We shall consider the most general case of a stochastic quantizer $q(y_N|y)$—the probability of a response y belonging to an abstract class y_N. A deterministic quantizer $f : Y \to Y_N$ is a special case in which q takes values zero or one only. In both cases, stimulus, response and reproduction form a Markov chain $X \to Y \to Y_N$. The quality of a quantization is characterized by a distortion function (appendix A.5.3). We shall look for a minimum distortion quantization using an information distortion function and discuss its relationship to the codebook estimation problem.

The quantization idea has been used implicitly in neuroscience for some time now. For example, the rate coding scheme effectively uses a deterministic quantizer to assign the neural response to classes based on the number of spikes that each pattern has. The metric space approach [40] uses an explicit cost (distortion) function to make different sequences identical if their difference according to the cost function is below a certain threshold. The cost function and identification threshold induce a deterministic quantization of the input space to a smaller output space. We decided to state the problem explicitly in the language of information theory, so that we could use the powerful methods developed in this context for putting all these ideas in a unified framework.

3.2. A distortion measure based on the mutual information

In engineering applications, the distortion function is usually chosen in a fairly arbitrary fashion [6, 13], and is the one that introduces structures in the original space, to be preserved by the quantization. We can avoid this arbitrariness, since we expect that the neural system is already reflecting pertinent structures of the sensory stimuli that we would like to preserve in the reproduction. Thus our choice of distortion function is determined by the informativeness of the quantization. The mutual information $I(X; Y)$ tells us how many different states on average can be distinguished in X by observing Y. If we quantize Y to Y_N (a reproduction with N elements), we can estimate $I(X; Y_N)$—the mutual information between X and the reproduction Y_N. Our information preservation criterion will then require that we choose a quantizer that preserves as much of the mutual information as possible, i.e. choose the quantizer $q(Y_N|Y)$ which minimizes the difference

$$D_I(Y, Y_N) = I(X; Y) - I(X; Y_N) \tag{1}$$

(note that $D_I \geqslant 0$). We use the functional D_I as a measure of the average distortion of the quality of a quantization. It can be interpreted as an *information distortion measure* (appendix A.5.3, appendix B), hence the symbol D_I. The only term that depends on the quantization is $I(X; Y_N)$ so we can reformulate the problem as the maximization of the effective functional $D_{\text{eff}} = I(X; Y_N)$. A closely related method using this cost function was recently presented in [37].

For several reasons it is useful to consider the full functional. First, we may choose to quantize the stimulus, in which case the quantizer is $q(x_N|x)$, or quantize both stimulus and response, in which case there are two quantizers. In these versions other parts of the information distortion are relevant. A second reason is that the average distortion can be rewritten as the expectation of a pointwise distortion function of a rather interesting form. Using the definition of the mutual information and the Markov relation $X \to Y \to Y_N$ between the spaces, we can express D_I (appendix B) as the expectation

$$D_I = E_{p(y,y_N)} d(y, y_N) \tag{2}$$

where

$$d(y, y_N) \equiv \mathrm{KL}(q(x|y)||q(x|y_N)) \tag{3}$$

is the Kullback–Leibler (KL) directed divergence of the input stimulus conditioned on a response y relative to the stimulus conditioned on a reproduction y_N. Intuitively, this measures how similar the stimulus partition induced by the quantization is to the partition induced by the sensory system. This expression is very appealing from a theoretical standpoint, due to various properties of the KL divergence. For example it allows us to describe the cases where the solution to the quantization problem will be an exact representation of the coding scheme. Indeed, because of the properties of KL [6], a reproduction with zero distortion is achieved if and only if $q(x|y) = q(x|y_N)$ almost everywhere, which is exactly the case with the coding scheme described in section 2.4.

3.3. Implementations

Using a quantization (deterministic or stochastic) of the output space (appendix A.5.2, [14]) allows us to control the exponential growth of required data. With this approach we estimate a quantity which is known to be a lower bound of the actual mutual information. We obtain a biased estimate but control the precision with which it can be estimated. Theorems from quantization theory (appendix A.5.2, [14]) insure that estimates of the quantized information quantities are always bounded by the original quantities and that a refinement of the quantization does not lower these estimates. In such an environment it is beneficial to fix the coarseness of the quantization (the size of the reproduction, N) and look for a quantization that minimizes the information distortion measure $D_I = I(X; Y) - I(X; Y_N)$ described previously.

3.3.1. Maximum entropy with nonlinear information distortion constraint. The problem of optimal quantization was formulated for a class of linear distortion functions [28] as a maximum-entropy problem [16]. We cannot use this analysis directly, since in our case the distortion function depends nonlinearly on the quantizer. However, we can still use the maximum-entropy formulation. The reasoning behind this is that, among all quantizers that satisfy a given set of constraints, the maximum-entropy quantizer does not implicitly introduce further restrictions in the problem. In this framework, the minimum-distortion problem is posed as a maximum-quantization-entropy problem with a distortion constraint:

$$\begin{array}{ll} \max\limits_{q(y_N|y)} H(Y_N|Y) & \text{constrained by} \\ D_I(q(y_N|y)) \leqslant D_\mathrm{o} & \text{and} \\ \sum\limits_{y_N} q(y_N|y) = 1 & \forall y \in Y. \end{array} \tag{4}$$

This is an ordinary constrained optimization problem that can be solved numerically with standard optimization tools. The cost function $H(Y_N|Y)$ is concave in $q(y_N|y)$, and the probability constraints $\sum_{y_N} q(y_N|y) = 1$ are linear in $q(y_N|y)$ [6]. The constraint D_I is also concave in $q(y_N|y)$ (theorem appendix B.1), which makes the whole problem one of concave maximization.

The problem with this formulation is that it relies on knowing D_I, which depends on the mutual information between X and Y. We can easily avoid the need for that by using the effective distortion $D_{\mathrm{eff}} \equiv I(X; Y_N)$. In this case, the optimization problem is

$$\begin{array}{ll} \max\limits_{q(y_N|y)} H(Y_N|Y) & \text{constrained by} \\ D_{\mathrm{eff}} \equiv I(q(y_N|y)) \geqslant I_\mathrm{o} & \text{and} \\ \sum\limits_{y_N} q(y_N|y) = 1 & \forall y \in Y. \end{array} \tag{5}$$

The solution to the optimization problem (5) depends on a single parameter I_{o}, which can be interpreted as the informativeness of the quantization. If $I_{\mathrm{o}} \leqslant 0$, the distortion constraint is always satisfied and we obtain only the unconstrained maximum entropy solution $q(y_N|y) = 1/N$ for all pairs (y, y_N). For $I_{\mathrm{o}} \geqslant 0$ the distortion constraint becomes active and the uniform quantizer is no longer a solution to the optimization problem. Because of the convexity of D_{eff}, the optimal solution will lie on the boundary of the constraint and thus carry $I(X; Y_N) = I_{\mathrm{o}}$ bits of information. Thus this formulation has a nice intuitive interpretation: 'find the maximum-entropy quantizer of Y which carries at least I_{o} bits of information about X'.

We can continue pushing I_{o} up for more informative solutions (with lower distortion) until we reach a point where the problem has no solutions. $I_{\mathrm{o}}^{\max}$ at this point is the best lower bound of $I(X; Y)$ for the N-element reproduction. Since the solution is continuous with respect to I_{o}, values near $I_{\mathrm{o}}^{\max}$ are also good lower bounds of $I(X; Y)$. By choosing one of the optimal quantizers near $I_{\mathrm{o}}^{\max}$, we can achieve the minimum distortion quantization.

3.3.2. Maximum cost with linear probability constraint. A standard approach to constrained optimization problems is through the use of Lagrange multipliers. The system (4) can be solved as the unconstrained optimization of

$$\max_{q(y_N|y)} \left(H(Y_N|Y) - \beta D_I(q(y_N|y)) + \sum_y \lambda_y \sum_{y_N} q(y_N|y) \right).$$

The solution depends on the parameters $(\beta, \{\lambda_y\})$ which can be found from the constraints

$$\begin{aligned} &D_I(q(y_N|y)) \leqslant D_{\mathrm{o}} \\ &\sum_{y_N} q(y_N|y) = 1 \qquad \forall y \in Y. \end{aligned}$$

Since β is a function of D_{o}, which is a free parameter, we can discard D_{o} and reformulate the optimization problem as finding the maximum of the cost function

$$\begin{aligned} \max_{q(y_N|y)} F(q(y_N|y)) \equiv {} & \max_{q(y_N|y)} \big(H(Y_N|Y) - \beta D_I(q(y_N|y)) \big) \\ & \text{constrained by} \\ & \sum_{y_N} q(y_N|y) = 1 \qquad \forall y \in Y. \end{aligned} \tag{6}$$

As in equation (5), we shall continue the discussion using the effective distortion D_{eff}. In this case,

$$\begin{aligned} \max_{q(y_N|y)} F(q(y_N|y)) \equiv {} & \max_{q(y_N|y)} \big(H(Y_N|Y) + \beta D_{\mathrm{eff}}(q(y_N|y)) \big) \\ & \text{constrained by} \\ & \sum_{y_N} q(y_N|y) = 1 \qquad \forall y \in Y. \end{aligned} \tag{7}$$

Even though this formulation is identical to (4), by transferring the nonlinear constraint in the cost function we can analyse the problem further. Following [28], we consider the behaviour of the cost function F at two limiting cases of β. When $\beta \to 0$, $F \to H(Y_N|Y)$ and the optimal solution is the unconstrained maximum-entropy solution $q(y_N|y) = 1/N$. This corresponds to the case $I_{\mathrm{o}} \leqslant 0$ in section 3.3.1. At the other limit, when $\beta \to \infty$, $F \to \beta D_{\mathrm{eff}}$ and the solution to the optimization problem approaches a maximum-D_{eff} (minimal-information-distortion D_I) solution. This is identical to the case where $I_{\mathrm{o}} \to I_{\mathrm{o}}^{\max}$ above. In order to avoid the divergence of the cost function with β, we rescale F to $F/(\beta + 1)$, which has the same extrema, but is bounded. Intermediate values of β produce intermediate solutions which connect the two limiting cases through a series of bifurcations. In [28] the parameter β was given the meaning of an annealing parameter and the whole procedure for a general class of distortion function was named 'deterministic annealing', drawing an analogy from a physical annealing process.

3.3.3. An implicit solution for the optimal quantizer. Further analysis of the problem uses the simplicity of the linear constraint in (7). Extrema of F can be found by setting its derivatives with respect to the quantizer $q(y_N|y)$ to zero. In the subsequent steps we shall explicitly use the assumption that all spaces are finite and discrete. The results for continuous random variables can easily be adapted from this using analogous methods from the calculus of variations. We use Latin indices (i, j, k) to denote members in the original spaces X, Y and Greek indices (μ, ν, η) for elements of the reproduction Y_N. With this in mind (appendix B.3), we continue to solve the Lagrange multiplier problem

$$\begin{aligned} 0 &= \left(\nabla\left(F + \sum_j \lambda_j \sum_\nu q(y_\nu|y_j)\right)\right)_{\nu k} \\ &= (\nabla H)_{\nu k} + \beta(\nabla D_{\text{eff}})_{\nu k} + \lambda_k \\ &= -p(y_k)\big(\ln q(y_\nu|y_k) + 1\big) + \beta(\nabla D_{\text{eff}})_{\nu k} + \lambda_k \\ \Leftrightarrow\ 0 &= \ln q(y_\nu|y_k) - \beta\frac{(\nabla D_{\text{eff}})_{\nu k}}{p(y_k)} - \mu_k \end{aligned} \tag{8}$$

where $\mu_k = \frac{\lambda_k}{p(y_k)} - 1$. Using this,

$$\begin{aligned} \ln q(y_\nu|y_k) &= \beta\frac{(\nabla D_{\text{eff}})_{\nu k}}{p(y_k)} + \mu_k \\ \Leftrightarrow\ q(y_\nu|y_k) &= \mathrm{e}^{\mu_k}\mathrm{e}^{\beta\left(\frac{(\nabla D_{\text{eff}})_{\nu k}}{p(y_k)}\right)}. \end{aligned} \tag{9}$$

The constraint on q requires that

$$\begin{aligned} 1 &= \sum_\nu q(y_\nu|y_k) \\ \Rightarrow 1 &= \mathrm{e}^{\mu_k}\sum_\nu \mathrm{e}^{\beta\left(\frac{(\nabla D_{\text{eff}})_{\nu k}}{p(y_k)}\right)} \\ \Leftrightarrow \mathrm{e}^{\mu_k} &= \frac{1}{\sum_\nu \mathrm{e}^{\beta\left(\frac{(\nabla D_{\text{eff}})_{\nu k}}{p(y_k)}\right)}}. \end{aligned} \tag{10}$$

We can substitute this in equation (9) and obtain an implicit expression for the optimal $q(y_\nu|y_k)$,

$$q(y_\nu|y_k) = \frac{\mathrm{e}^{\beta\left(\frac{(\nabla D_{\text{eff}})_{\nu k}}{p(y_k)}\right)}}{\sum_\nu \mathrm{e}^{\beta\left(\frac{(\nabla D_{\text{eff}})_{\nu k}}{p(y_k)}\right)}}. \tag{11}$$

Note that ∇D_{eff} is a function of the quantizer $q(y_N|y)$. As $\nabla D_{\text{eff}} = -\nabla D_I$, the implicit solution in terms of the information distortion D_I is

$$q(y_N|y) = \frac{\mathrm{e}^{-\beta\frac{\nabla D_I}{p(y)}}}{\sum_{y_N}\mathrm{e}^{-\beta\frac{\nabla D_I}{p(y)}}}. \tag{12}$$

In practice, the expression (11) can be iterated for a fixed value of β to obtain a solution for the optimization problem, starting from a particular initial state. For small β, before the first bifurcation described in [28] occurs, the obvious initial condition is the uniform solution $q(y_N|y) = 1/N$. The solution for one value of β can be used as the initial condition for a subsequent value of β because solutions are continuous with respect to β.

3.4. Approximations to the codebook

The optimal information distortion procedure can help us resolve the neural decoding problem we outlined in section 2.4. Indeed, in the limit of no distortion (no information loss), an identity transformation preserves all the structure of the output, but usually is unavailable due to lack of data. When we quantize the neural response by fixing the size of the reproduction to N, this bounds our estimate of D_{eff} to be no more than $\log N$ bits. In the ideal case $\max D_{\text{eff}} \equiv \max I(X; Y_N) \approx \log N$, but in general it will be lower. On the other hand we have a bound $I(X; Y_N) \leqslant I(X; Y)$. Since $\log N$ increases with N and $I(X; Y)$ is a constant, these two independent bounds intersect for some $N = N_{\text{c}}$, at which point adding more elements to Y_N does not improve the distortion measure. If $I(X, Y_N)$ increases with N until $N = N_{\text{c}}$ and then levels off, we can identify the correct N_{c} by examining the behaviour of the expected distortion (or, equivalently, $D_{\text{eff}} \equiv I(X; Y_N)$) as a function of N, given sufficient data. We take the elements of Y_N as the labels of the equivalence classes which we wanted to find. The quantizer $q(y_N|y)$ gives the probability of a response y belonging to an equivalence class y_N. Rose [28] conjectured that the optimal quantizer for low distortions (high β) is deterministic (or effectively deterministic, in the case of duplicate classes). In this case the responses associated with class y_N are $\mathcal{Y}_N = \{y|q(y_N|y) \approx 1\}$. The optimal quantizer also induces a coding scheme from $X \to Y_N$ by $p(y_N|x) = \sum_y q(y_N|y)p(y|x)$. This is the most informative approximation of the original relation $p(x|y)$. It induces the quantization $X \to X_N$ by associating x_N with the stimulus set $\mathcal{X}_N = \{x|p(y_N|x) \approx 1\}$ of all x which correspond to the same output class y_N. The resulting relation $p(y_N|x_N)$ is almost bijective and so we recover an almost complete reproduction of the model described in section 2.3.

If there are not enough data to support a complete recovery even under the reduced data requirements, the algorithm has to stop earlier. The criterion we use in such a case is that the estimate of D_{eff} does not change with N *within its error bounds* (obtained analytically or by statistical re-estimation methods like bootstrap, or jack-knife). Then $N < N_{\text{c}}$ and the quantized mutual information is at most $\log N$. We can recover at most N classes and some of the original classes will be combined. The quantizer may also not be deterministic due to lack of enough data to resolve uncertainties. Thus we can recover a somewhat impoverished picture of the actual input/output relationship, which can be refined automatically as more data become available, by increasing N and repeating the optimization procedure.

4. Results

We shall discuss the application of the method described so far to a few synthetic test cases. Applying it to physiological data from a sensory system involves additional difficulties associated with the estimates of D_I for complex input stimuli, which are dealt with elsewhere [10, 11].

4.1. Random clusters

We present the analysis of the probability distribution shown in figure 4(*a*). In this model we assume that X represents a range of possible stimulus properties and Y represents a range of possible spike train patterns. We have constructed four clusters of pairs in the stimulus/response

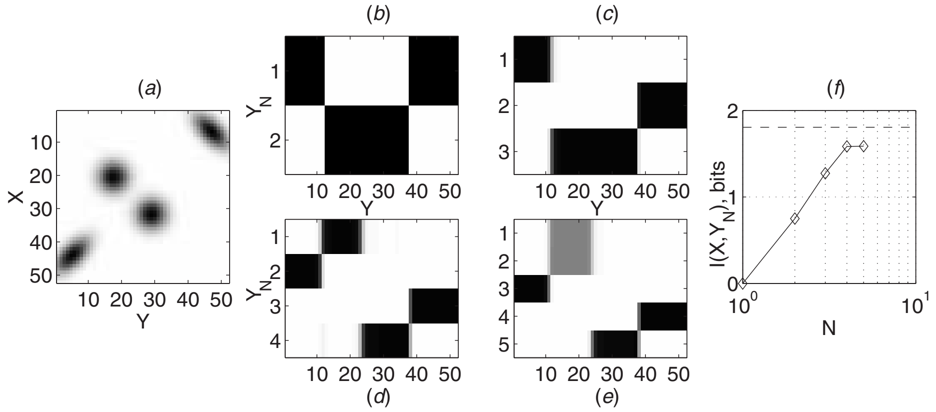

Figure 4. (*a*) A joint probability for the relation between two random variables X and Y with 52 elements each. (*b*)–(*e*) The optimal quantizers $q(y_N|y)$ for different numbers of classes. These panels represent the conditional probability $q(y_N|y)$ of a pattern y from (*a*) (horizontal axis) belonging to class y_N (vertical axis). White represents zero, black represents one and intermediate values are represented by levels of grey. The behaviour of the mutual information with increasing N can be seen in the log–linear plot (*f*). The dashed curve is $I(X;Y)$, which is the least upper bound of $I(X;Y_N)$.

space. Each cluster corresponds to a range of responses elicited by a range of stimuli. This model was chosen to resemble the model of coding and decoding with jointly typical sequences (section 2.4). The mutual information between the two sequences is about 1.8 bits, which is comparable to the mutual information conveyed by single neurons about stimulus parameters in several unrelated biological sensory systems [9, 18, 25, 33]. For this analysis we assume the original relation between X and Y is known (the joint probability $p(x, y)$ is used explicitly).

The results of the application of our approach are shown in panels (*b*)–(*f*) of figure 4. The grey-scale map in these, and later, representations of the quantizer depicts zero with white, one with black and intermediate values with levels of grey. When a two-class reproduction is forced (*b*), the algorithm recovers an incomplete representation. The representation is improved for the three-class refinement (*c*). The next refinement (*d*) separates all the classes correctly and recovers most of the mutual information. Further refinements (*e*) fail to split the classes and are effectively identical to (*d*) (note that classes 1 and 2 in (*e*) are almost evenly populated and the class membership there is close to a uniform one-half). The quantized mutual information (*f*) increases with the number of classes approximately as $\log N$ until it recovers about 90% of the original mutual information ($N = 4$), at which point it levels off.

Further details of the course of the optimization procedure that lead to the optimal quantizer in panel (*d*) are presented in figure 5. The behaviour of D_{eff} as a function of the annealing parameter β can be seen in the top panel. Snapshots of the optimal quantizers for different values of β are presented on the bottom row (panels 1–6). We can observe the bifurcations of the optimal solution (1–5) and the corresponding transitions of the effective distortion. The abrupt transitions (1 → 2, 2 → 3) are similar to the ones described in [28] for a linear distortion function. We also observe transitions (4 → 5) which appear to be smooth in D_{eff} even though the solution for the optimal quantizer undergoes a bifurcation.

A random permutation of the rows and columns of the joint probability in figure 4(*a*) has the same channel structure. The quantization is identical to the case presented in figure 4 after applying the inverse permutation and fully recovers the permuted classes (i.e., the quantization is contravariant with respect to the action of the permutation group).

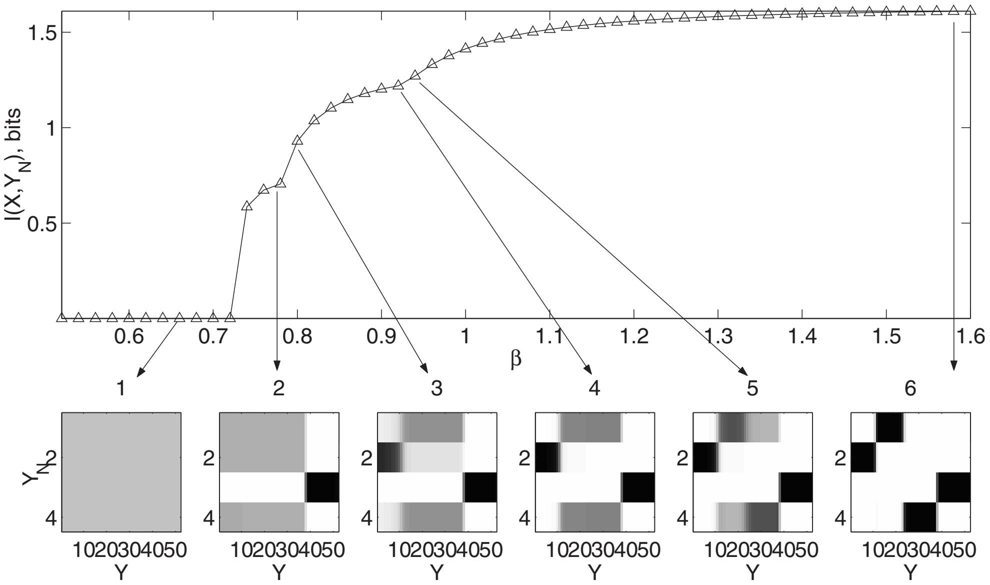

Figure 5. Behaviour of D_{eff} (top) and the optimal quantizer $q(y_N|y)$ (bottom) as a function of the annealing parameter β.

4.2. Hamming code

There exist noise-correcting codes that transform blocks of symbols at the input to larger blocks of binary symbols in such a way that the original block can be recovered even after some noise perturbs the codewords [6]. For this example we are going to use a simple noise-correcting code—the Hamming (7, 4) code (H74). It operates on binary blocks of size four and expands each block to seven binary symbols by adding three parity bits. Blocks can be considered as vectors in a Boolean vector space. The input $x \in \mathcal{Z}_2^4$ and the output $y \in \mathcal{Z}_2^7$. The Hamming code then can be described as a linear operator on these spaces:

$$y = H^T \cdot x$$

where

$$H = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}$$

and any addition is modulo 2. This code can detect and correct one error (a flipped bit) in a block of seven.

The properties of H74 are known very well. We are going to use this system as another testbed to evaluate the performance of different implementations of the algorithms described earlier. In this case we used data sampled from the joint probability $p(X, Y)$ to perform all estimates. The two variables X and Y are related in the following manner: we generated a sequence of points from a uniform source in $\mathcal{Z}_2^4$ (about 10^4 for this example) and applied H74 to it. Two independent copies of the result were perturbed by noise, which flipped a random bit of each sample. The two perturbed outputs are X_o and Y_o. The relation between these is shown in figure 6. There are four bits of mutual information between X_o and Y_o. A permutation of the rows and columns, which orders equivalent sequences of H74, makes the relation much easier to comprehend (figure 7(a)). The variables X and Y are permuted versions of X_o and

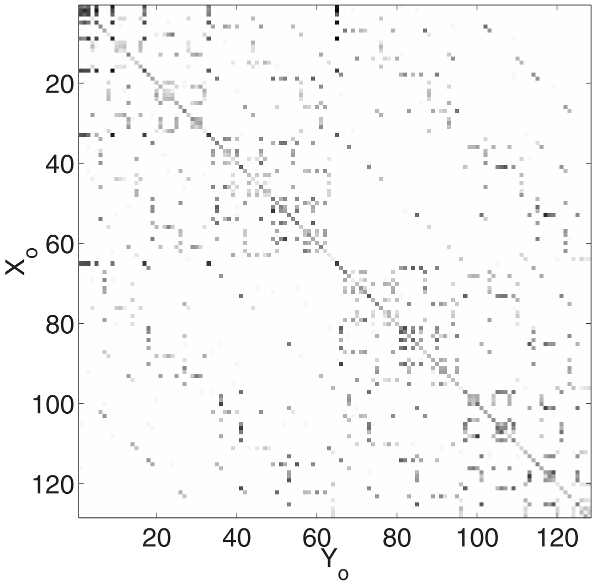

Figure 6. The joint probability between X_o and Y_o in the original H74 variables.

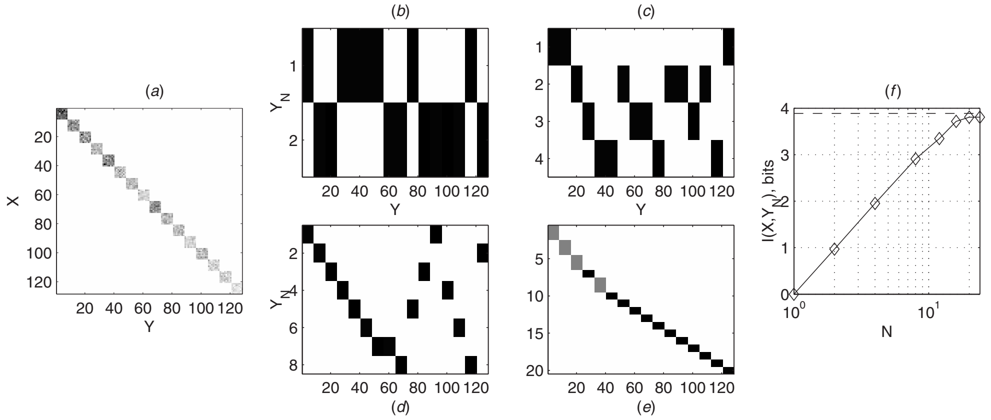

Figure 7. The joint probability between X and Y after permuting rows and columns (a) and optimal quantizations for different number of classes (b)–(e). The behaviour of the mutual information with increasing N can be seen in the log–linear plot (f). The dashed curve is $I(X;Y)$.

Y_o and have the same relation between them. As mentioned in section 4.1, any permutation of variables leaves the information distortion invariant. Thus the results for (X, Y) and (X_o, Y_o) are equivalent.

The results can be seen in figure 7. The reproduction classes were permuted to follow roughly the relation 7(a). As in section 4.1, the algorithm recovers several incomplete representations (b)–(e), each one a refinement of the previous. Refinements beyond the correct number of clusters (16) fail to improve the distortion by much (e), and some of the classes are effectively copies of each other (e.g. 1 and 2). The quantized mutual information (f) increases with the number of classes approximately as $\log N$ until it recovers almost all of the original mutual information ($N = 16$), at which point it levels off.

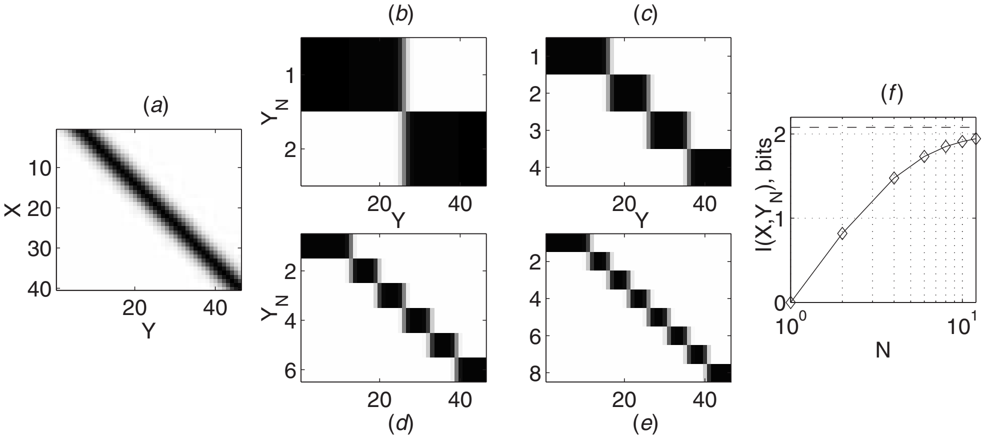

Figure 8. A joint probability for a linear relation between two random variables X and Y (*a*) with optimal quantization (*b*)–(*e*) for different number of classes. The behaviour of the mutual information with increasing N can be seen in (*f*). The dashed curve is $I(X; Y)$.

4.3. Linear encoding

We also applied the algorithm to a case which, unlike the previous cases, does not have clearly defined clusters. This model tries to simulate the process of a physical measurement where X represents a physical system and Y is the set of possible results from a measurement. In this example we model a linear relation between X and Y and Gaussian measurement noise, that is

$$Y = kX + \eta$$

where $\eta \in \mathcal{N}(0, \sigma)$ is drawn from a normal distribution with zero mean and variance σ^2. The particular relation we used (figure 8(*a*)) contains about two bits of mutual information.

The results can be seen in figure 8. The algorithm recovers a series of representations (*b*)–(*e*), where each quantizer is a refinement of the previous one. The reproduction classes were permuted to roughly follow the original linear relation. There is no natural stopping point and so the quantized mutual information $I(X; Y_N)$ approaches the original mutual information $I(X; Y)$ gradually. This is in contrast to the previous two cases, where the rate of change of $I(X; Y_N)$ abruptly decreased after some N.

The course of the bifurcations here also differs from the previous cases (figure 9). Again, the reproduction classes were permuted to roughly follow the original linear relation. There are no obvious abrupt transitions and the uncertainty of the quantizer is resolved smoothly with β.

5. Conclusions and discussion

This paper has two goals. The first goal is to formulate a precise model of an early stage of a sensory system as a communication channel, and describe the properties of this model. In general, a communication channel is fully described by the conditional probability of response given a stimulus. Using ideas from information theory on optimal information transmission in the presence of noise, and the method of jointly typical sequences, we have demonstrated the existence of equivalence classes of stimulus/response pairs, which we have called *codeword classes*. A coding scheme in this system can be described by an almost deterministic map

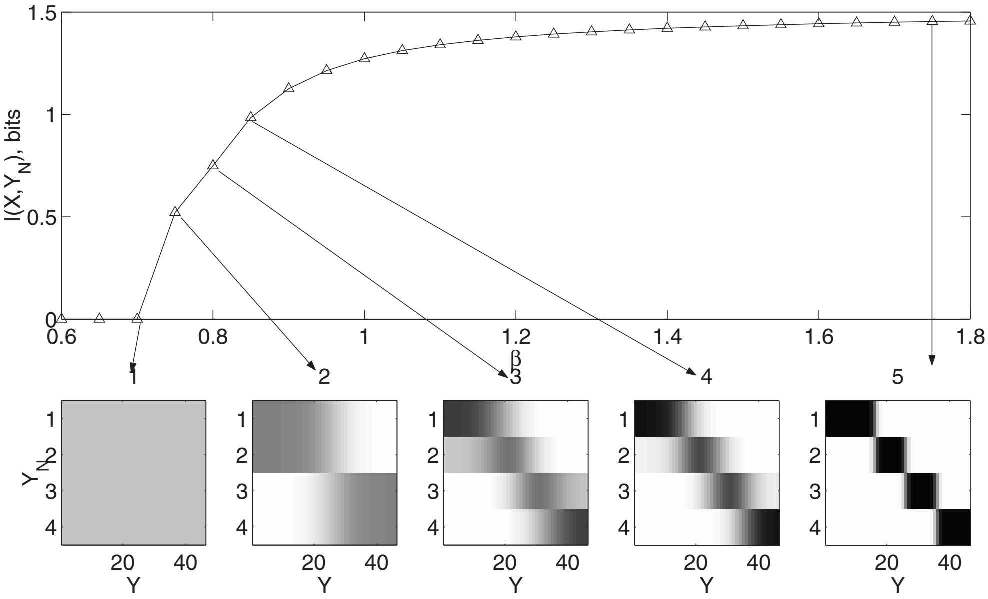

Figure 9. Behaviour of D_{eff} (top) and the optimal quantizer $q(y_N|y)$ (bottom) as a function of the annealing parameter for the linear encoding case.

when restricted to the codeword classes. The number of codeword classes is related to the mutual information $I(X; Y)$ between stimulus and response.

The second goal of this paper is to provide a method for recovering the structure of such a model from observations. Characterizing the relation between individual elements in the stimulus and response spaces has been shown to require large numbers of data points, increasing exponentially with the length of spike sequences (T) and the number of neurons (N) considered. We choose to recover an impoverished description of the coding scheme by quantizing the responses to a reproduction set of a few elements. To assess the quality of the reproduction, we define the information distortion $D_I = I(X; Y) - I(X; Y_N)$, which measures how much information is lost in the quantization process. For a fixed reproduction size N we pose the optimization problem of finding the quantization with smallest distortion, as the one which preserves most of the information present in the original relation between X and Y. Refining the reproduction by increasing N was shown to decrease the distortion. We demonstrate empirically on a set of synthetic problems that, if the original relation contains almost disjoint clusters, a sufficiently fine optimal quantization recovers them completely. If the quantization is too coarse, then some of the clusters will be combined, but in such a way that a large fraction of the original information is still preserved.

In realistic cases of physiological recordings, there are not usually enough data to support a sufficiently fine quantization. In such cases, we are forced to accept a coarse quantization which does not recover all the structure of a particular neural coding scheme. The criterion we have adopted for stopping the refinement process is when the estimate of the information distortion does not change within its error bounds, which may be obtained analytically or by statistical re-estimation procedures.

Many of the coding schemes currently in use can be seen as special cases of the method we present here. A rate code can be described as a deterministic quantization to the set of integers. The quantizer assigns all spike patterns with the same number of spikes to the same equivalence class. A spike latency code can be seen as a quantization to classes determined by the latency and jitter of the spike's timing. A stimulus feature is decoded as in the rate code case, based on which latency 'class' a spike falls in. The metric space approach [40] uses an explicit cost (distortion) function to determine what different sequences are identical: they are equivalent if according to the cost function their difference is below a certain threshold. The cost function and identification threshold induce a deterministic quantization of the input space to a smaller output space.

Most current approaches to studying neural coding rely on formulating a hypothesis about the coding scheme a neural system may use and then using observations to estimate parameters of the hypothesis. The complexity of the hypothesis determines the amount of data needed for reliable estimates of the necessary parameters. The method presented in this report offers a means for data-driven hypothesis formulation. When we stop the refinement of the reproduction due to lack of data, we effectively formulate a hypothesis about the most informative coding scheme that can be supported with the available amount of data. When more observations become available, the hypothesis can be refined automatically to include them for a better approximation.

Recent applications of information theory to problems of neural coding [5, 7, 33] concentrate on estimating information-theoretic quantities without regard to the actual stimulus/response relationship that produced them. The strong point in that approach—that it does not require a detailed understanding of the coding process—is also a weak point from another perspective. As we mentioned, without the model of a coding scheme we present here, many candidate hypotheses must be investigated one by one. For example, in [5] the basic elements are a pair of events, in [4] a combination of rate and latency codes. Using the approach described here, we can characterize coding schemes without reference to the underlying physical processes that produce them. It could be that the codeword classes that emerge after the analysis are well described by a simple mechanism, but we also have the ability to analyse and describe quite succinctly coding schemes which have not been considered so far.

We developed the information distortion method from a purely practical necessity—the need to describe a neural coding scheme on large input and output spaces. Our earlier research [9, 35] and others [23, 25, 33, 41] estimate just a few bits of mutual information per neuron in several distinct sensory systems. In view of our model of a coding scheme this suggests that there are just a few codeword classes that need to be identified, regardless of the size of the response space. Thus it was natural to devise a method that clustered the neural representation in a few large sets, while preserving most of the mutual information. We were therefore quite excited to find another recently developed approach [37], which suggested a very similar method ('the information bottleneck') for completely abstract reasons (the authors attempt to extract 'meaningful' or 'relevant' information from a pair of interacting systems). The different motivations are obvious. We explicitly concentrate on the case of finite (albeit large) spaces, so that the method is applicable to computer-recorded data and numerical analysis. The 'information bottleneck' method [37] uses a variational approach to continuous random variables, which is better suited for abstract analysis. We cannot assess its applicability to actual stimulus/response datasets, since no examples of its performance are presented in [37]. The only application we found cited there [32] suffers from two unfortunate choices. First, it applies the method to a problem which, albeit real, is not well understood, so we could not distinguish the limitations of the algorithm from the constraints of the problem itself. Second, it chooses to forgo the simplicity offered by the small reproduction space and probabilistic clustering [28] and instead uses an *ad hoc* deterministic clustering method to find an approximation of the solution to their problem, which makes assessing the properties of the method even more difficult. We hope that further developments in both approaches will eventually converge to a unified method for data analysis.

The information distortion D_I seems to be extremely well suited for uncovering the structure of information channels. While appendix B addresses some of the properties of this object, more research is needed to elucidate these further. Since it is concave and its first and second derivatives are continuous except at the boundaries of its domain, it is amenable to various analytical techniques, which can help clarify the structure of the optimal solution and bifurcations at different values of the distortion. However, this is beyond the scope of this paper.

It is interesting to note that, although we had neural coding in mind while developing the information distortion method, the ensuing analysis is in no way limited to nervous systems. Indeed, the constraints on the pair of signals we analyse are so general that they can represent *almost any* pair of interacting physical systems. In this case, finding a minimal information distortion reproduction allows us to recover certain aspects of the interaction between the two physical systems, which may improve considerably any subsequent analysis performed on them. It is also possible to analyse parts of the structure of a single physical system Y, if X is a system with known properties (e.g. a signal generator, controlled by a researcher) and is used to perturb Y. These cases point to the exciting possibility of obtaining a more automated approach for succinct descriptions of arbitrary physical systems through the use of minimal information distortion quantizers.

Acknowledgments

This research was supported in part by NIH grants MH12159 (AGD) and MH57179 (JPM). The authors are indebted to Penio Penev (Rockefeller University), Tomáš Gedeon, Zane Aldworth, Zuzana Gedeon, Kay Kirkpatrick (Montana State University) and the two anonymous reviewers for numerous discussions and suggestions which helped shape this paper.

Appendix A. Information theory

The goal of this section is to summarize results from information theory, pertinent to the body of the paper. Statements of standard results are made with some precision so that the interested reader can trace the foundation of our discussion about coding and quantization earlier in the paper. An extended treatment of these topics and more can be found in [6]. A more formal treatment can be found in [14], where the subject is approached with greater care and mathematical precision.

Appendix A.1. Source space, probability measure and random variables

An *information source* is a mathematical model for a physical system that produces a succession of symbols in a manner which is unknown to us and is treated as random. The set $\mathcal{X}$, containing all possible output symbols, is called the *alphabet* of the source. Let A be a σ-field of subsets of $\mathcal{X}$. The treatment of events (sets of sequences of symbols) is achieved through assigning a probability measure $p(x) = \Pr\{X = x\}, x \in \mathcal{X}$ to $(\mathcal{X}, A)$. The triplet $(\mathcal{X}, A, p)$ is often called a *random variable*, X, when the context of the alphabet and measure is clear.

In cases with two or more alphabets, one can define a probability measure $p(x, y)$ on the product space $(x, y) \in \mathcal{X} \times \mathcal{Y}$. In this context the induced probability measures on the individual spaces are called *marginal measures* or *marginals*. An important measure which emerges from this setup is the *conditional probability*, defined as

$$p(x|y) \equiv \frac{p(x, y)}{p(y)}, \qquad \forall(x, y) \in \mathcal{X} \times \mathcal{Y}.$$

Two random variables are called *independent* if

$$p(x|y) = p(x) \qquad \forall y \in \mathcal{Y}.$$

A *measurable function* defined on $(\mathcal{X}, A)$ and taking values in another measurable space $(\mathcal{Y}, B)$ is a mapping $f : \mathcal{X} \to \mathcal{Y}$ with the property that

$$F \in B \Longrightarrow f^{-1}(F) = \{y : f(y) \in F\} \in A.$$

Whenever it is well defined (usually when the measurable function is in a linear vector space, e.g. R^n), one can perform integration, which is a functional on the space of measurable functions. In probability theory this is called the *expectation* of the measurable function $g(X)$ and is defined as

$$E_p g(X) \equiv \sum_{x \in \mathcal{X}} g(x) p(x)$$

where the sum is replaced by an integral for continuous alphabets.

Appendix A.2. Information-theoretic quantities

The basic concepts of information theory are entropy and mutual information. The notion of *mutual information* is introduced as an integral measure of the degree of dependence between a pair of variables. The concept of *entropy* can then be understood as the *self-information* of a random variable. They are both special cases of a more general integral quantity called *relative entropy* (or KL divergence [20]), which is an integral measure of the difference between two probability distributions. While mutual information is well defined for both discrete and continuous alphabets, entropy for continuous alphabets is a problematic quantity, undefined for many measures of interest. Fortunately many identities and bounds on mutual information are still valid if one uses another integral measure from probability—the KL divergence, or relative entropy, between two probability measures on the same event space:

$$\mathrm{KL}(p\|q) = E_p \log\left(\frac{p(x)}{q(x)}\right) \tag{A.1}$$

where p and q are two different probability measures on $\mathcal{X}$. $\mathrm{KL}(p||q)$ quantifies the difference between two probability measures on the same sample space and is extensively used in probabilistic decision theory. Note that, as with most of the quantities here, KL depends only on probability measures and not on the elements of the space, which can be non-numeric. Thus expectations are always well defined here irrespective of the structure of the event space. The usual definitions of these quantities use a base two logarithm, so any further references to the log function implicitly assume base two. On rare occasions we shall use the natural logarithm, denoted by the symbol ln.

Appendix A.2.1. Mutual information. In order to measure the statistical independence between two random variables X and Y it is useful to introduce the notion of *mutual information*. It is defined as the KL distance between the joint probability $p(x, y)$ on the product alphabet $\mathcal{X} \times \mathcal{Y}$ and the product of marginal probabilities $p(x)$ and $p(y)$. $I(X; Y)$ is equal to zero if and only if X and Y are independent:

$$I(X; Y) \equiv \mathrm{KL}\,(p(x, y), p(x)p(y)) = E_{p(x,y)} \log \frac{p(x, y)}{p(x)p(y)}. \tag{A.2}$$

The mutual information is symmetric with respect to its arguments: $I(X; Y) = I(Y; X)$.

Appendix A.2.2. Entropy. A measure of self-information of a probability distribution is given by the *entropy* $H(X)$:

$$H(X) \equiv E_p \log \frac{1}{p(x)} = I(X;X). \tag{A.3}$$

In communication theory H measures the average information that each symbol carries when sampled.

The *joint entropy* of a pair of random variables is defined as

$$H(X,Y) \equiv E_{p(x,y)} \log \frac{1}{p(x,y)}. \tag{A.4}$$

We also define the *conditional entropy* of a random variable given another as the expected value of the entropies of the conditional distributions:

$$H(X|Y) \equiv E_{p(x)} H(Y|X=x) = E_{p(x,y)} \log \frac{1}{p(y|x)}. \tag{A.5}$$

Appendix A.2.3. Information identities. There are various identities connecting I and H [6]. Here is a short list of the most frequently used.

$$\begin{aligned}
&H(X) = I(X;X)\\
&I(X;Y) = H(X) - H(X|Y)\\
&I(X;Y) = H(Y) - H(Y|X)\\
&I(X;Y) = H(X) + H(Y) - H(X,Y)\\
&H(X_1,\ldots,X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1},\ldots,X_n)\\
&I(X_1,\ldots,X_n;Y) = \sum_{i=1}^{n} I(X_i;Y|X_{i-1},\ldots,X_1).
\end{aligned} \tag{A.6}$$

Appendix A.3. Typical sequences

Theorem appendix A.1 (Asymptotic equipartition property (AEP)). *If $X_1, X_2, \ldots X_n$ are independent and identically distributed random variables (i.i.d.) with probability measure $p(x)$, then*

$$n^{-1} \log p(X_1, X_2, \ldots, X_n)^{-1} \to H(X) \tag{A.7}$$

in probability.

The theorem is a simple consequence of the weak law of large numbers. For its proof and that of most other theorems consult [6]. Its extension to arbitrary ergodic finite-valued processes is known as the *Shannon–McMillan–Breiman theorem.* All definitions and theorems in this section will be presented in their i.i.d. form, but in general they are correct for ergodic sources. The AEP allows us to define a structure in the set of events $\mathcal{X}$.

Definition. *The typical set A_ϵ^n with respect to $p(x)$ is the set of sequences $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ with the following property:*

$$2^{-n(H(X)+\epsilon)} \leqslant p(x_1, x_2, \ldots, x_n) \leqslant 2^{-n(H(X)-\epsilon)}. \tag{A.8}$$

The typical set has the following properties.

Theorem appendix A.2 (Properties of typical sequences).

(i) *If* $(x_1, x_2, \ldots, x_n) \in A^n_\epsilon$, *then* $|n^{-1} \log p(x_1, x_2, \ldots, x_n)^{-1} - H(X)| \leqslant \epsilon$.
(ii) $\Pr\{A^n_\epsilon\} > 1 - \epsilon$ *for n sufficiently large.*
(iii) $(1 - \epsilon)2^{n(H(X)-\epsilon)} \leqslant |A^n_\epsilon| \leqslant 2^{n(H(X)+\epsilon)}$ *for n sufficiently large. Here* $|A|$ *is the number of elements in set A.*

Appendix A.4. Jointly typical sequences

When analysing information channels, we deal with two sets of random sequences—input and output. In this case it is necessary to consider the combined behaviour of the pair (X, Y).

Definition. *The set* A^n_ϵ *of jointly typical sequences* $\{(x^n, y^n)\}$ *with respect to the distribution* $p(x, y)$ *is the set*

$$A^n_\epsilon = \{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : |-n^{-1} \log p(x^n) - H(X)| < \epsilon,$$
$$|-n^{-1} \log p(y^n) - H(Y)| < \epsilon, |-n^{-1} \log p(x^n, y^n) - H(X, Y)| < \epsilon\}, \tag{A.9}$$

where $p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i)$.

Theorem appendix A.3 (Properties of jointly typical sequences). *Let* (X^n, Y^n) *be i.i.d. pair sequences of length n. Then the following hold.*

(i) $\Pr(A^n_\epsilon) > 1 - \epsilon$.
(ii) $|A^n_\epsilon| \leqslant 2^{n(H(X,Y)+\epsilon)}$ *for n sufficiently large.*
(iii) *If* $(\tilde{X}^n, \tilde{Y}^n)$ *are a pair of random variables with joint probability* $p(x^n, y^n) = p(x^n)p(y^n)$ *(i.e.* $\tilde{X}^n$ *and* $\tilde{Y}^n$ *are independent with the same marginal distributions as* X^n *and* Y^n*), then for sufficiently large n*

$$(1 - \epsilon)2^{-n(I(X,Y)+3\epsilon)} \leqslant \Pr\left((\tilde{X}^n, \tilde{Y}^n) \in A^n_\epsilon\right) \leqslant 2^{-n(I(X,Y)-3\epsilon)}.$$

The properties of jointly typical sequences can be used to prove one of the principal theorems in information theory: the channel coding theorem. For lack of space we cannot do this here. Instead we just state this important result. A complete proof using these techniques can be found in [6].

Definition. *A discrete channel is a system consisting of an input source X, an output source Y and a transition probability* $p(y|x)$ *of observing the output symbol y given x was sent. The channel is memoryless if the transition probability is conditionally independent of previous channel inputs and outputs.*

Definition. *The channel capacity of a discrete memoryless channel is*

$$C = \max_{p(x)} I(X; Y). \tag{A.10}$$

Theorem appendix A.4 (The channel coding theorem). *For any channel with capacity C and every rate* $R < C$ *there exists a sequence of codes with this rate and maximum probability of error* $\lambda^{(n)} \rightarrow 0$.

Conversely, any sequence of codes with $\lambda^{(n)} \rightarrow 0$ *must have* $R \leqslant C$.

Appendix A.5. Distortion theory

Appendix A.5.1. The data processing inequality. The data processing inequality is used in the proofs of most of the statements later, so we shall take some time to give proper definitions and proofs.

Definition. *Random variables X, Y, Z form a Markov chain (denoted by $X \to Y \to Z$) if the conditional distribution of Z depends only on Y and is conditionally independent of X. In this case, the joint probability can be written as*

$$p(x, y, z) = p(z|y)p(y|x)p(x). \tag{A.11}$$

In particular, if $z = f(y)$, where f is a deterministic function of y, then $X \to Y \to Z$. Also note that if $X \to Y \to Z$ then $Z \to Y \to X$ as well.

Theorem appendix A.5 (Data processing inequality). *If $X \to Y \to Z$, then $I(X; Z) \leqslant I(X; Y)$.*

Proof. By the chain rule (last item in (A.6)) we can expand the mutual information in two different ways.

$$\begin{aligned} I(X; Y, Z) &= I(X; Z) + I(X; Y|Z) \\ &= I(X; Y) + I(X; Z|Y). \end{aligned}$$

Since X and Z are independent given Y, this implies $I(X; Z|Y) = 0$. On the other hand, $I(X; Y|Z) \geqslant 0$, so

$$I(X; Z) \leqslant I(X; Y).$$

In particular, if $Z = f(Y)$ then $I(X; Z) \leqslant I(X; Y)$. □

One intepretation of this inequality is that any physical measurement, deterministic or not, will usually *decrease* the information carried by one random variable about another.

Appendix A.5.2. Quantization. A random variable Y can be related to another random variable Y_N through the process of *quantization* [6, 14]. Y_N is referred to as the *reproduction* of Y. Quantizers can be deterministic (functions) or stochastic (given through a conditional probability). The size of the reproduction space is smaller than the size of the quantized space.

A *deterministic quantizer* (or just quantizer, also referred to as hard clustering [28]) is any simple measurable function $f : X \to X_f$ from X to a reproduction space X_f with finitely many elements x^i_f. The quantizer f induces a partition $\{Q^i_f\}$ such that

(i) $Q^i_f \subset X$,
(ii) $Q^i_f \bigcap Q^j_f = \emptyset$ if $i \neq j$,
(iii) $\bigcup_i Q^i_f = X$.

The information quantities in the reproduction X_f are

$$H(X_f) = E_{p(x_f)} \log \frac{1}{p(x_f)} = E_{p(Q_f)} \log \frac{1}{p(Q_f)} \tag{A.12}$$

$$I(X_f, Y_g) = E_{p(x_f, y_g)} \log \frac{p(x_f, y_g)}{p(x_f)p(y_g)} = E_{p(Q_f, Q_g)} \log \frac{p(Q_f, Q_g)}{p(Q_f)p(Q_g)}. \tag{A.13}$$

The quantizer h refines f ($h > f$) if the partition Q_h of X induced by h refines the partition Q_f induced by f. Q_h refines Q_f ($Q_h > Q_f$) if any $Q^i_h \in Q_h$ is a subset of some

$Q_f^j \in Q_f$. Note that in this case $X \to X_h \to X_f$; i.e., the reproductions form a certain Markov chain with the original space.

A *stochastic quantizer* (soft clustering [28]) is a mapping q from one probability measure space X to another X_q. The mapping is given by the conditional probability $q(x_q|x)$, which can be interpreted as the probability of x belonging to the reproduction class x_q. The source space induces a probability measure on the reproduction space by $p(x_q) = \sum_x q(x_q|x)p(x)$. A stochastic quantizer can also be considered as a communication channel. The deterministic quantizer discussed above is a special case of a stochastic quantizer with $q(x_q|x) = 1$ if $x \in x_q$ and zero otherwise.

The information quantities in X_q are

$$H(X_q) = E_{p(x_q)} \log \frac{1}{p(x_q)} \tag{A.14}$$

$$I(X_q, Y_g) = E_{p(x_q, y_g)} \log \frac{p(x_q, y_g)}{p(x_q)p(y_g)}. \tag{A.15}$$

Unlike the deterministic case, here we do not have a nice inverse image of the quantizer map to define refinements in the usual way. Thus we choose a different property of the quantizer as a definition of refinement: the quantizer h refines f ($h > f$) if $X \to X_h \to X_f$ (that is, X_f is upstream of X_h in a Markov chain).

There are some properties of quantizers (deterministic or stochastic) that are useful for discussing information transmission.

If $h > f$, then from the Markov relation and theorem appendix A.5 it follows that

$$\begin{aligned} &H(Y|X_h) \leqslant H(Y|X_f) \\ &I(Y, X_h) \geqslant I(Y, X_f). \end{aligned} \tag{A.16}$$

For any X (discrete or continuous), and any quantizer f

$$I(Y, X) \geqslant I(Y, X_f) \tag{A.17}$$

that is, estimates in the quantized spaces are always lower bounds of the actual information quantities.

When X is continuous, $H(X_f)$ diverges with refinements [14]. $I(X, Y)$ on the other hand can always be obtained as the least upper bound over all refinements. Without further constraints, this is achieved by a deterministic quantizer [28].

The statements above suggest that quantizations could provide lower-bound estimates of H and I with controlled precision, since the size of the pattern set is fixed by the size of the quantized space and could be potentially much lower than the size of the original pattern space. This allows us to obtain more precise estimates of the quantities in question.

Appendix A.5.3. Distortion theory. The quality of a quantization is the topic of distortion theory [6]. It is characterized by a distortion function $d : X \times X_q \to \mathcal{R}^+$. The distortion function $d(x, x_q)$ measures how well x is represented by x_q. In general it can be arbitrary. The *expected distortion*

$$D(q(x_q|x)) = E_{q(x_q|x)p(x)} d(x, x_q)$$

measures the quality of quantization by the quantizer $q(x_q|x)$.

Definition. *The rate distortion function $R(D)$ for a source X, reproduction X_q and distortion $d(x, x_q)$ is defined [6] as*

$$R(D) = \min_{q(x_q|x):D(p(x,x_q))\leqslant D} I(X; X_q). \tag{A.18}$$

The minimization here is over all stochastic quantizers $q(x_p|x)$ which satisfy the distortion constraint.

The solution to (A.18) is a standard minimization problem of the convex function $I(X, X_q)$ over the convex set $\{q(x_q|x) : q(x_q|x) \geqslant 0; \sum_{x_q} q(x_q|x) = 1; D(q(x_q|x)p(x)) \equiv \sum_{x,x_p} q(x_p|x)p(x)\, \mathrm{d}(x, x_p) \leqslant D\}$. Analytically it can be solved by using Lagrange multipliers to minimize

$$J(q(x_q|x)) = I(q(x_q|x)p(x)) + \beta D(q(x_q|x)) + \sum_x \mu(x) \sum_{x_q} q(x_q|x)$$

and find λ and $\mu(x)$ from the constraints.

Appendix B. Properties of the information distortion function

Appendix B.1. Definition

The (average) information distortion D_I between a random variable Y and its reproduction Y_N is defined through another random variable, X, related to Y so that $X \to Y \to Y_N$ form a Markov chain. In this case, we define the *information distortion* between Y and Y_N as

$$D_I(Y, Y_N) = I(X, Y) - I(X, Y_N). \tag{B.1}$$

The only part that depends on the quantizer $q(y_N|y)$ is $I(X, Y_N)$ so we can concentrate on the properties of $D_{\mathrm{eff}} = I(X, Y_N)$.

Appendix B.2. Properties

D_I is a bounded function of the quantizer. Indeed, since $D_I = I(X, Y) - I(X, Y_N)$ and $0 \leqslant I(X, Y_N) \leqslant I(X, Y)$, this implies $0 \leqslant D_I \leqslant I(X, Y)$.

D_I is an *expected* distortion—an integral characteristic of the relation between the whole sets Y and Y_N. We can write it in a form that includes explicitly the expectation of a pointwise distortion function. Indeed, using that $p(x, y) = \sum_{y_N} p(x, y, y_N)$ and $p(x, y_N) = \sum_y p(x, y, y_N)$, we have

$$\begin{aligned}
D_I &= I(X, Y) - I(X, Y_N) \\
&= \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} - \sum_{x,y_N} p(x, y_N) \log \frac{p(x, y_N)}{p(x)p(y_N)} && \text{(B.2)} \\
&= \sum_{x,y,y_N} p(x, y, y_N)\big(\log p(x|y) - \log p(x|y_N)\big) && \text{(B.3)} \\
&= \sum_{y,y_N} p(y, y_N) \sum_x p(x|y) \log \frac{p(x|y)}{p(x|y_N)} \\
&= \sum_{y,y_N} p(y, y_N)\, \mathrm{KL}\big(p(x|y)\| p(x|y_N)\big). && \text{(B.4)}
\end{aligned}$$

Here (B.3) uses the Bayes property $p(x, y)/p(y) = p(x|y)$ and $\log p(x)$ is common for the two parts and cancels. Step (B.4) uses the Markov property $p(x, y, y_N) = p(x|y)p(y, y_N)$. This shows that the information distortion

$$D_I = E_{p(y,y_N)} \mathrm{KL}\big(p(x|y)\| p(x|y_N)\big) \tag{B.5}$$

is the expectation of the KL divergence (A.1) of $p(x|y)$ with respect to $p(x|y_N)$. Unlike the pointwise distortion functions usually investigated in information theory [6, 28], this one depends on the quantizer $q(y_N|y)$, thorough $p(x|y_N)$.

In the process of investigating the information distortion function we may have to evaluate it at values of $q(y_N|y)$ which are *not* conditional probabilities. The only requirement we want to impose on q is that it be non-negative: $q(y_N|y) \geqslant 0$. For this reason we need to define certain relations which are automatically true when $q(y_N|y)$ is a conditional probability. We shall use that $p(x, y)$ is a probability function. With that in mind,

$$\begin{aligned} p(x) &= \sum_y p(x, y) \\ p(y) &= \sum_x p(x, y). \end{aligned} \tag{B.6}$$

We *define*

$$p(x, y_N) \equiv \sum_y q(y_N|y) p(x, y) \tag{B.7}$$

$$\begin{aligned} p(y_N) &\equiv \sum_x p(x, y_N) = \sum_y q(y_N|y) \sum_x p(x, y) \\ q(y_N|x) &\equiv \sum_y q(y_N|y) p(y|x). \end{aligned} \tag{B.8}$$

All these are automatically true when $q(y_N|y)$ is a conditional probability. If $q(y_N|y)$ is *not* a conditional probability, none of the above are probabilities but all are non-negative. Also $p(x) \neq \sum_{y_N} p(x, y_N)$ unless $q(y_N|y)$ is a conditional probability. It is still the case that

$$p(x, y_N) = q(y_N|x) p(x) \tag{B.9}$$

(the Bayes property) because of (B.6) and the definitions (B.7). With these definitions, the natural extensions for the information quantities (A.2), (A.5) for arbitrary non-negative $q(y_N|y)$ are

$$\begin{aligned} I(X, Y_N) &\equiv \sum_{x, y_N} p(x, y_N) \log \frac{p(x, y_N)}{p(x) p(y_N)} \\ H(Y_N|Y) &\equiv \sum_{y, y_N} q(y_N|y) p(y) \log q(y_N|y). \end{aligned} \tag{B.10}$$

Lemma appendix B.1 (Convexity of D_I). *The information distortion function D_I is a concave function of any $q(y_N|y) \geqslant 0$.*

Proof. Consider $D_I = I(X, Y) - I(X, Y_N)$. The first term is constant with respect to the $q(y_N|y)$, so it is sufficient to consider only the second term. Using the definition of I above (B.10) and (B.9) we can see that $I(X, Y_N)$ is a function of $q(y_N|x)p(x)$. For a fixed $p(x)$, $I(q(y_N|y))$ is a convex function of $q(y_N|x)$ [6] (shown there for $q(y_N|x)$ a probability, but easily extensible for any $q(y_N|x) \geqslant 0$). Hence,

$$I(\lambda q_1(y_N|x) + (1 - \lambda) q_2(y_N|x)) \leqslant \lambda I(q_1(y_N|x)) + (1 - \lambda) I(q_2(y_N|x)) \tag{B.11}$$

for any $q_1, q_2 \geqslant 0$. We need to show that I is also a convex function of $q(y_N|y)$. Because of (B.7),

$$q(y_N|x) = \sum_y q(y_N|y) p(y|x). \tag{B.12}$$

Consider $\tilde{I}(q(y_N|y)) \equiv I\big(\sum_y q(y_N|y) p(y|x)\big)$. For any $q_1(y_N|y), q_2(y_N|y) \geqslant 0$,

$$\tilde{I}(\lambda q_1(y_N|y) + (1 - \lambda) q_2(y_N|y)) = I\bigg(\sum_y \big(\lambda q_1(y_N|y) + (1 - \lambda) q_2(y_N|y)\big)\big) p(y|x)\bigg)$$

$$
\begin{aligned}
&= I\Big(\lambda \sum_y q_1(y_N|y)p(y|x) + (1-\lambda)\sum_y q_2(y_N|y)p(y|x)\Big) \\
&= I\big(\lambda q_1(y_N|x) + (1-\lambda)q_2(y_N|x)\big) \qquad \text{(B.13)} \\
&\leqslant \lambda I(q_1(y_N|x)) + (1-\lambda)I(q_2(y_N|x)) \\
&= \lambda I\Big(\sum_y q_1(y_N|y)p(y|x)\Big) + (1-\lambda)I\Big(\sum_y q_2(y_N|y)p(y|x)\Big) \\
&= \lambda \tilde{I}(q_1(y_N|y)) + (1-\lambda)\tilde{I}(q_2(y_N|y)) \qquad \text{(B.14)}
\end{aligned}
$$

where (B.13) follows from (B.12) and (B.14) is a consequence of the convexity of I (B.11). This finishes the proof that $I(X, Y_N)$ is a convex function of $q(y_N|y)$. Since $D_I \propto -I(X, Y_N) \Rightarrow D_I$ is a concave function of the quantizer $q(y_N|y)$. □

Appendix B.3. Derivatives

D_I is a function of the quantizer $q(y_N|y)$. The effective distortion $D_{\text{eff}} = I(X, Y_N)$ has the same derivatives with respect to the quantizer as D_I, with opposite sign, so we shall consider only the derivatives of D_{eff}. In this section we shall explicitly use the assumption that all spaces are finite and discrete. We do not use anywhere the fact that $q(y_N|y)$ is a conditional probability. The final results are thus ready for programming in a computer or for further finite-dimensional analysis. The results for continuous random variables can easily be adapted from here using analogous methods from calculus of variations. We use Latin indices (i, j, k) to denote members in the original spaces X, Y and Greek indices (μ, ν, η) for elements of the reproduction Y_N. While differentiating, we shall use the natural logarithm (ln) in all definitions and rescale the results to base two logarithm (log) at the end.

In terms of the quantizer,

$$D_{\text{eff}} = \sum_{i,\mu} p(x_i, y_\mu) \ln \frac{p(x_i, y_\mu)}{p(x_i)p(y_\mu)} \qquad \text{(B.15)}$$

where (B.7)

$$
\begin{aligned}
p(x_i, y_\mu) &\equiv \sum_j q(y_\mu|y_j)p(x_i, y_j) \\
p(y_\mu) &\equiv \sum_j q(y_\mu|y_j)p(y_j).
\end{aligned} \qquad \text{(B.16)}
$$

It is beneficial to calculate the derivatives of (B.16) before attempting this for D_{eff}. We use the fact that $\frac{\partial q(y_\mu|y_j)}{\partial q(y_\nu|y_k)} = \delta_{\mu\nu}\delta_{jk}$. The derivatives are

$$
\begin{aligned}
\frac{\partial p(x_i, y_\mu)}{\partial q(y_\nu|y_k)} &= \delta_{\mu\nu} p(x_i, y_k) \\
\frac{\partial p(y_\mu)}{\partial q(y_\nu|y_k)} &= \delta_{\mu\nu} p(y_k).
\end{aligned} \qquad \text{(B.17)}
$$

Using (B.17),

$$
\begin{aligned}
(\nabla D_{\text{eff}})_{\nu k} &\equiv \frac{\partial D_{\text{eff}}}{\partial q(y_\nu|y_k)} \\
&= \frac{\partial}{\partial q(y_\nu|y_k)} \sum_{i,\mu} p(x_i, y_\mu) \ln \frac{p(x_i, y_\mu)}{p(x_i)p(y_\mu)}
\end{aligned}
$$

$$= \sum_{i,\mu} \frac{\partial p(x_i, y_\mu)}{\partial q(y_\nu|y_k)} \ln \frac{p(x_i, y_\mu)}{p(x_i)p(y_\mu)} + p(x_i, y_\mu)\frac{\partial}{\partial q(y_\nu|y_k)}\big(\ln p(x_i, y_\mu) - \ln p(y_\mu)\big)$$

$$= \sum_{i,\mu} \delta_{\mu\nu} p(x_i, y_k) \ln \frac{p(x_i, y_\mu)}{p(x_i)p(y_\mu)} + \delta_{\mu\nu} p(x_i, y_\mu)\left(\frac{p(x_i, y_k)}{p(x_i, y_\mu)} - \frac{p(y_k)}{p(y_\mu)}\right)$$

$$= \sum_i p(x_i, y_k) \ln \frac{p(x_i, y_\mu)}{p(x_i)p(y_\mu)} - \overbrace{\sum_i p(x_i, y_k)}^{\equiv p(y_k)} + \frac{p(y_k)}{p(y_\mu)} \overbrace{\sum_i p(x_i, y_\mu)}^{\equiv p(y_\mu)} \qquad \text{(B.18)}$$

$$\Rightarrow (\nabla D_{\text{eff}})_{\nu k} = \sum_i p(x_i, y_k) \ln \frac{p(x_i, y_\nu)}{p(x_i)p(y_\nu)}. \qquad \text{(B.19)}$$

The second derivatives are

$$\frac{\partial^2 D_{\text{eff}}}{\partial q(y_\eta|y_l)\partial q(y_\nu|y_k)} = \frac{\partial}{\partial q(y_\eta|y_l)} \sum_i p(x_i, y_k) \ln \frac{p(x_i, y_\nu)}{p(x_i)p(y_\nu)}$$

$$= \sum_i p(x_i, y_k) \frac{\partial}{\partial q(y_\eta|y_l)}\big(\ln p(x_i, y_\nu) - \ln p(y_\nu)\big)$$

$$= \sum_i p(x_i, y_k)\delta_{\nu\eta}\left(\frac{p(x_i, y_l)}{p(x_i, y_\nu)} - \frac{p(y_l)}{p(y_\nu)}\right) \qquad \text{(B.20)}$$

$$\Rightarrow \frac{\partial^2 D_{\text{eff}}}{\partial q(y_\eta|y_l)\partial q(y_\nu|y_k)} = \delta_{\nu\eta}\left(\sum_i \frac{p(x_i, y_k)\, p(x_i, y_l)}{p(x_i, y_\nu)} - \frac{p(y_k)p(y_l)}{p(y_\nu)}\right). \qquad \text{(B.21)}$$

In all cases we assume that all relevant quantities are absolutely continuous with respect to one another, so that all divisions can be performed. In practice this means that any optimization has to be restricted away from zero, since the gradients diverge there. When a deterministic quantizer (hard clustering) is required, the optimization can be brought close to a boundary ($q(y_N|y) \approx 0$ for some (y_N, y)) and then thresholded to obtain the deterministic map.

When posing the optimization problem (section 3.3), we encounter the functional $F = H(Y_N|Y) - \beta D_I$. To analyse it further we also need the derivatives of $H(Y_N|Y)$:

$$(\nabla H)_{\nu k} \equiv \frac{\partial H(Y_N|Y)}{\partial q(y_\nu|y_k)}$$

$$= -\frac{\partial}{\partial q(y_\nu|y_k)} \sum_{j,\mu} q(y_\mu|y_j)p(y_j) \ln q(y_\mu|y_j)$$

$$= -\sum_{j,\mu} p(y_j)\delta_{\mu\nu}\delta_{jk}\big(\ln q(y_\mu|y_j) + 1\big)$$

$$\Rightarrow (\nabla H)_{\nu k} = -p(y_k)\big(\ln q(y_\nu|y_k) + 1\big). \qquad \text{(B.22)}$$

The second derivatives are

$$\frac{\partial^2 H(Y_N|Y)}{\partial q(y_\eta|y_l)\partial q(y_\nu|y_k)} = -\frac{\partial}{\partial q(y_\eta|y_l)} p(y_k)\big(\ln q(y_\nu|y_k) + 1\big)$$

$$= -\frac{p(y_k)}{q(y_\nu|y_k)}\delta_{\nu\eta}\delta_{kl}. \qquad \text{(B.23)}$$

To obtain the results when the information quantities are measured in bits, all of the derivatives above should be divided by $\ln 2 \approx 0.6931$.

References

[1] Adrian E D 1928 *The Basis of Sensation: the Action of the Sense Organs* (New York: Norton)

[2] Atick J J 1992 Could information theory provide an ecological theory of sensory processing? *Network: Comput. Neural Syst.* **3** 213–51

[3] Barlow H B 1961 Possible principles underlying the transformation of sensory messages *Sensory Communications* ed W A Rosenblith (Cambridge, MA: MIT Press)

[4] Berry M J and Meister M 2001 Firing events: fundamental symbols in the neural code of retinal ganglion cells? *Computational Neuroscience: Trends in Research* ed J Bower (Amsterdam: Elsevier)

[5] Brenner N, Strong S P, Koberle R, Bialek W and de Ruyter van Steveninck R R 2000 Synergy in a neural code *Neural Comput.* **12** 1531–52

[6] Cover T and Thomas J 1991 *Elements of Information Theory (Wiley Series in Communication)* (New York: Wiley)

[7] de Ruyter van Steveninck R R, Lewen G D, Strong S P, Koberle R and Bialek W 1997 Reproducibility and variability in neural spike trains *Science* **275** 1805–8

[8] Dimitrov A G 1998 Aspects of cortical information processing *PhD Thesis* The University of Chicago

[9] Dimitrov A G and Miller J P 2000 Natural time scales for neural encoding *Neurocomputing* **32–3** 1027–34

[10] Dimitrov A G, Miller J P and Aldworth Z 2002 Spike pattern-based coding schemes in the cricket cercal sensory system *Computational Neuroscience: Trends in Research* ed J Bower to be published

[11] Dimitrov A G, Miller J P, Aldworth Z and Gedeon T 2001 Non-uniform quantization of neural spike sequences through an information distortion measure *Computational Neuroscience: Trends in Research* ed J Bower (Amsterdam: Elsevier)

[12] Georgopoulos A P, Schwartz A B and Kettner R E 1986 Neuronal population coding of movement direction *Science* **233** 1416–9

[13] Gersho A and Gray R M 1992 *Vector Quantization and Signal Compression* (Dordrecht: Kluwer)

[14] Gray R M 1990 *Entropy and Information Theory* (Berlin: Springer)

[15] Hubel D and Wiesel T 1961 Receptive fields, binocular interaction and functional architecture in the cat's visual cortex *J. Physiol. (Lond.)* **195** 215–43

[16] Jaynes E T 1982 On the rationale of maximum-entropy methods *Proc. IEEE* **70** 939–52

[17] Johnson D H, Gruner C M, Baggerly K and Seshagiri C 2001 Information-theoretic analysis of the neural code *J. Comput. Neurosci.* **10** 47–70

[18] Kjaer T W, Hertz J A and Richmond B J 1994 Decoding cortical neuronal signals: network models, information estimation and spatial tuning *J. Comput. Neurosci.* **1** 109–39

[19] Koch C and Segev I (ed) 1992 *Methods in Neuronal Modeling* (Cambridge, MA: MIT Press)

[20] Kullback S 1959 *Information Theory and Statistics* (New York: Wiley)

[21] Meister M and Berry M J 1999 The neural code of the retina *Neuron* **22** 435–50

[22] Oram M W, Wiener M C, Lestienne R and Richmond B J 1999 The stochastic nature of precisely timed spike patterns in visual system neural responses *J. Neurophysiol.* **81** 3021–33

[23] Panzeri S, Schultz S R, Treves A and Rolls E T 1999 Correlations and the encoding of information in the nervous system *Proc. R. Soc.* B **266** 1001–12

[24] Penev P S and Sirovich L 2000 The global dimensionality of face space *Proc. 4th Int. Conf. on Automatic Face and Gesture Recognition (Grenoble, 2000)* (Los Alamitos, CA: IEEE Computer Society Press) pp 264–70

[25] Reinagel P and Reid R 2000 Temporal coding of visual information in the thalamus *J. Neurosci.* **20** 5392–400

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

[27] Rose K 1994 A mapping approach to rate distortion computation and analysis *IEEE Trans. Inform. Theory* **40** 1939–52

[28] Rose K 1998 Deteministic annealing for clustering, compression, classification, regression, and related optimization problems *Proc. IEEE* **86** 2210–39

[29] Salinas E and Abbott L F 1994 Vector reconstruction from firing rates *J. Comput. Neurosci.* **1** 89–107

[30] Schultz S R and Panzeri S 2001 Temporal correlations and neural spike train entropy *Phys. Rev. Lett.* **86** 5823–6

[31] Shannon C E 1948 A mathematical theory of communication *Bell Syst. Tech. J.* **27** 623–56

[32] Slonim N and Tishby N 2000 Agglomerative information bottleneck *Advances in Neural Information Processing Systems* vol 12, ed S A Solla, T K Leen and K-R Müller (Cambridge, MA: MIT Press) pp 617–23

[33] Strong S P, Koberle R, de Ruyter van Steveninck R R and Bialek W 1998 Entropy and information in neural spike trains *Phys. Rev. Lett.* **80** 197–200

[34] Theunissen F and Miller J P 1995 Temporal encoding in nervous systems: a rigorous definition *J. Comput. Neurosci.* **2** 149–62

[35] Theunissen F, Roddey J C, Stufflebeam S, Clague H and Miller J P 1996 Information theoretic analysis of dynamical encoding by four primary sensory interneurons in the cricket cercal system *J. Neurophysiol.* **75** 1345–59

[36] Theunissen F E and Miller J P 1991 Representation of sensory information in the cricket cercal sensory system. II. Information theoretic calculation of system accuracy and optimal tuning curve widths of four primary interneurons *J. Neurophysiol.* **66** 1690–703

[37] Tishby N, Pereira F C and Bialek W 2000 The information bottleneck method *LANL Preprint* http://arXiv.org/abs/physics/0004057

[38] Treves A and Panzeri S 1995 The upward bias in measures of information derived from limited data samples *Neural Comput.* **7** 399–407

[39] Victor J D 2000 How the brain uses time to represent and process visual information *Brain Res.* **886** 33–46

[40] Victor J D and Purpura K 1997 Metric-space analysis of spike trains: theory, algorithms, and application *Network: Comput. Neural Syst.* **8** 127–64

[41] Warland D 1991 Reading between the spikes: real-time processing in neural systems *PhD Thesis* University of California at Berkeley