Neural Coding: The Efficient Coding Hypothesis

Angus Chadwick

School of Informatics, University of Edinburgh, UK

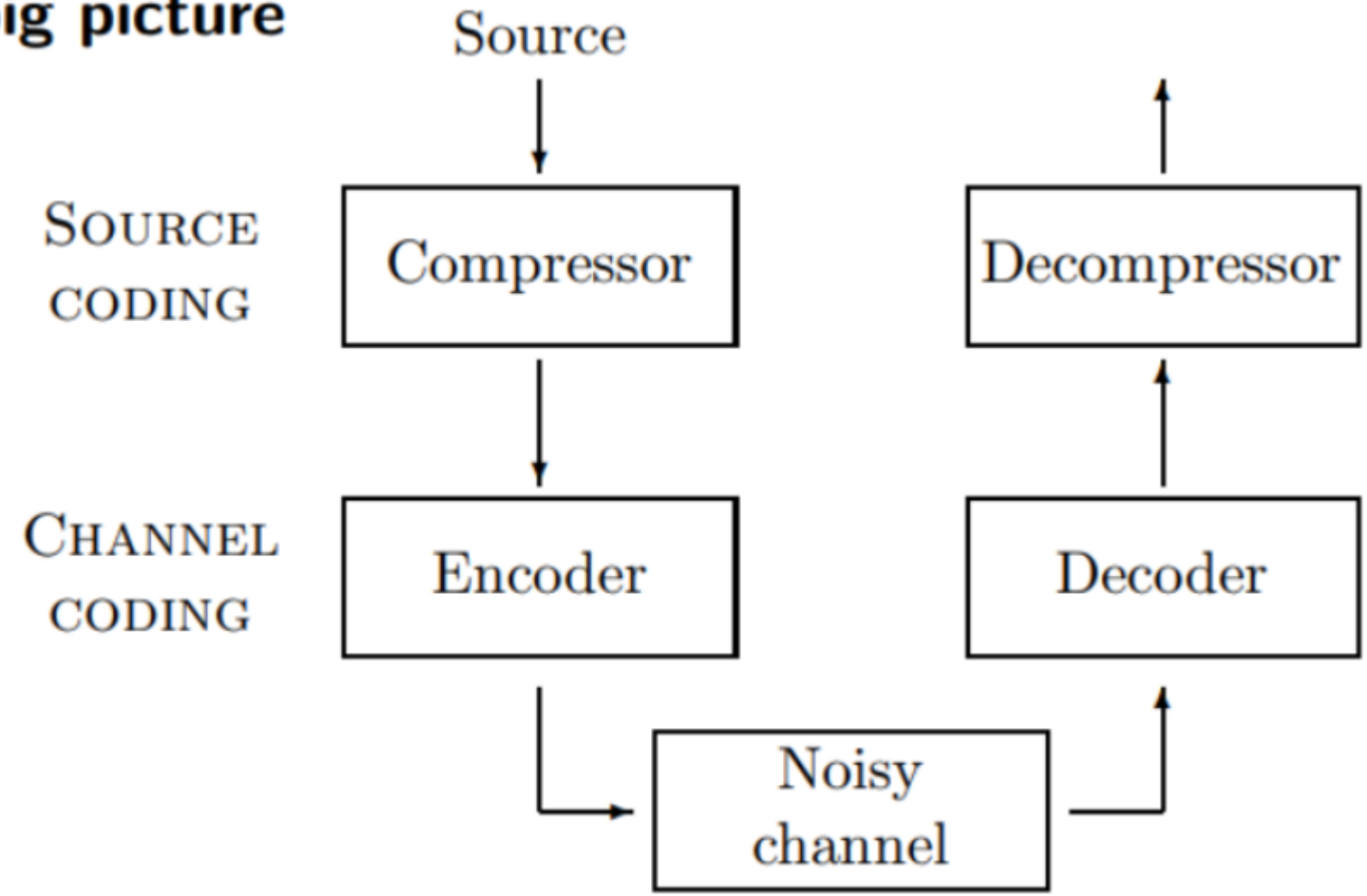
Computational Neuroscience (Lecture 8, 2024/2025)

Outline of Lecture

- Introduction to neural coding
- Information theory (surprise, entropy, and mutual information)
- The data processing inequality
- The efficient coding hypothesis
- Applications: filtering in the retina, histogram equalisation in the blowfly visual system

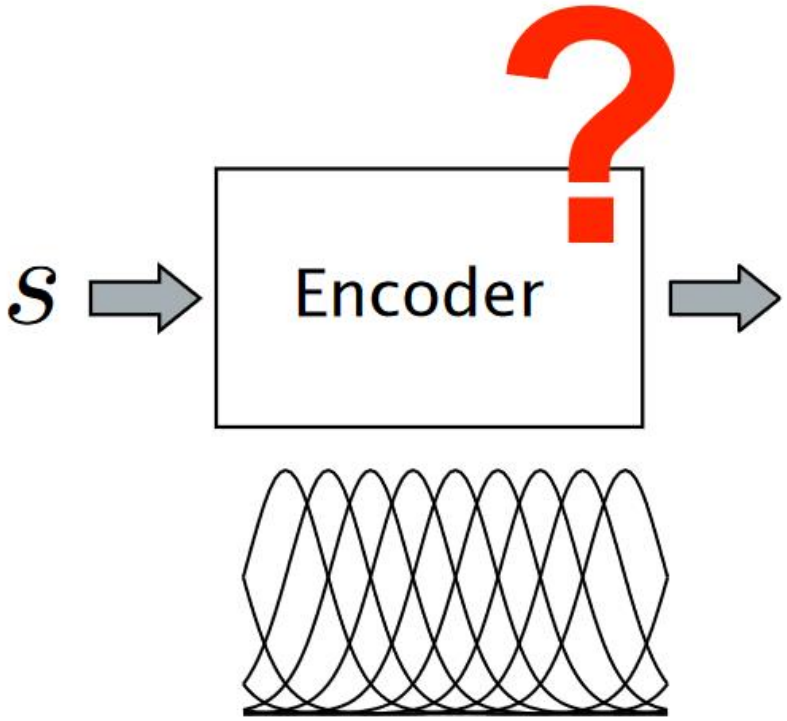
Neural Coding

The big picture

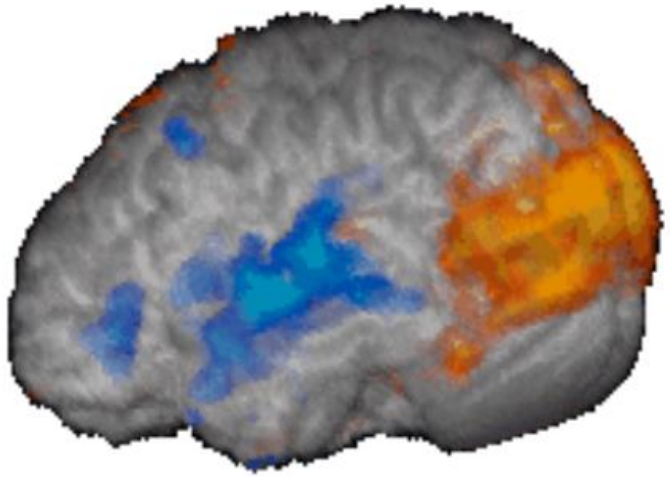


Neural Coding: Encoding vs Decoding

The World



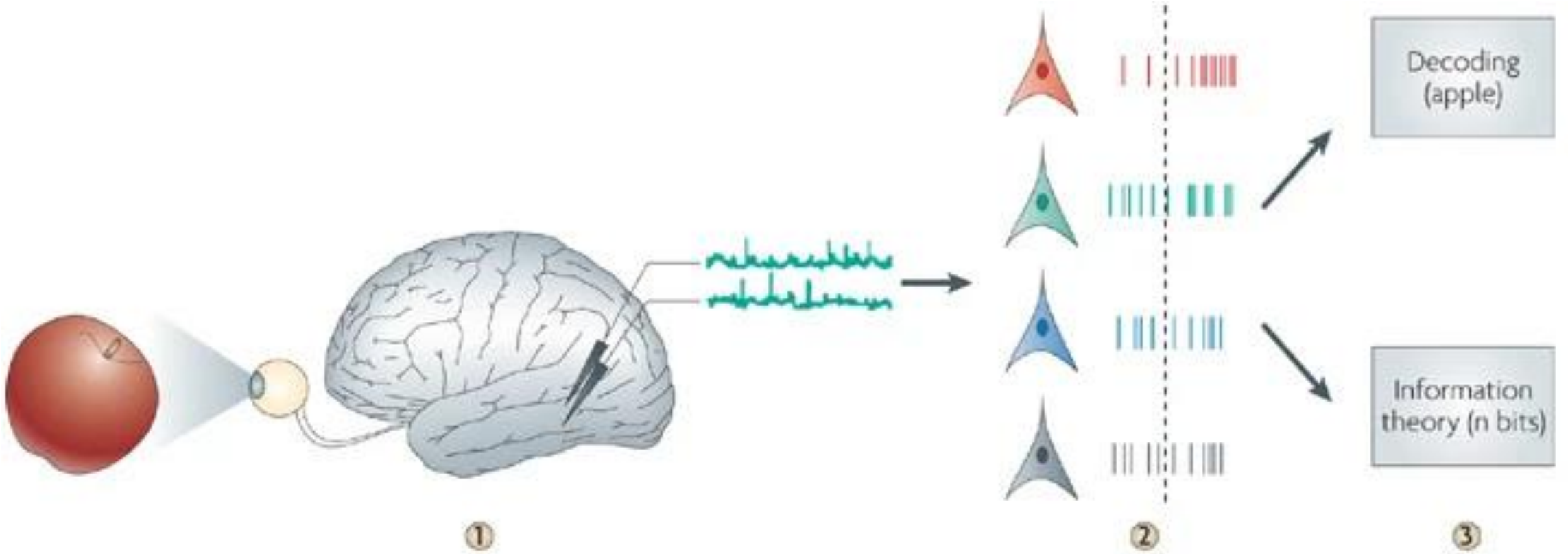
Activity in the brain



Encoding: P(Brain|World)

Decoding: P(World|Brain)

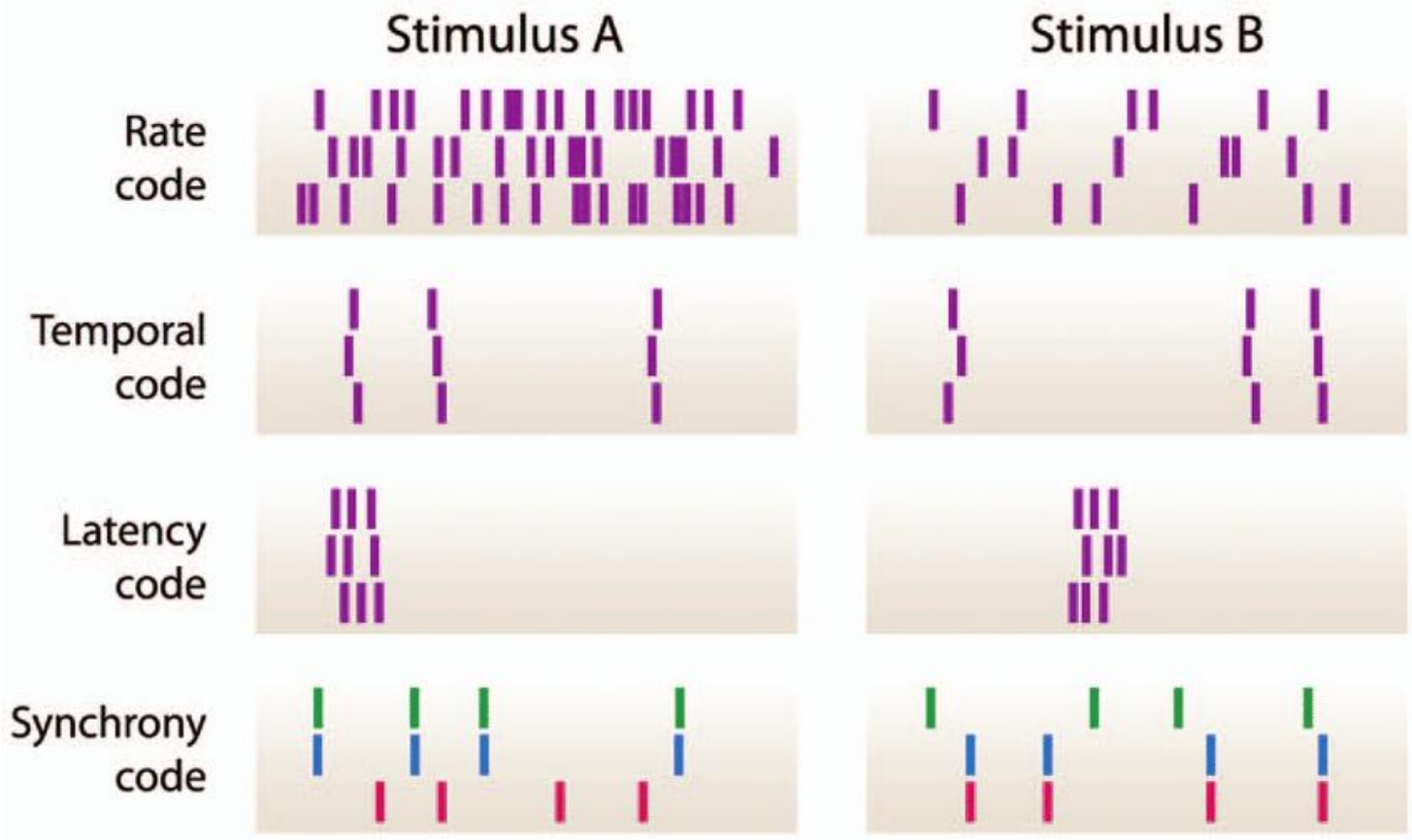
Neural Coding: Decoding vs Information Theory



Nature Reviews | Neuroscience

Neural Coding: Spike Timing vs Rate

How is information represented by neurons? Number of spikes? Spike times? Relationships between spikes of different neurons?



Information Theory

Developed by Shannon for his masters thesis (!)

Answers question such as:

- What is the optimal code for sending messages down a noisy channel?
- How can signals be compressed to transmit them more efficiently?
- What are the fundamental limits at which signals can be encoded, transmitted, and decoded?

The Bell System Technical Journal

Vol. XXVII *July, 1948* *No. 3*

A Mathematical Theory of Communication

By C. E. SHANNON

Introduction

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.

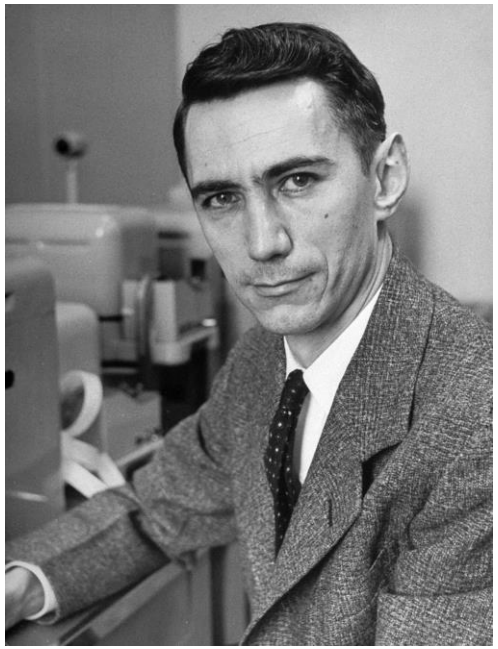
If the number of messages in the set is finite then this number or any monotonic function of this number can be regarded as a measure of the information produced when one message is chosen from the set, all choices being equally likely. As was pointed out by Hartley the most natural choice is the logarithmic function. Although this definition must be generalized considerably when we consider the influence of the statistics of the message and when we have a continuous range of messages, we will in all cases use an essentially logarithmic measure.

The logarithmic measure is more convenient for various reasons:

1. It is practically more useful. Parameters of engineering importance

[1] Nyquist, H., "Certain Factors Affecting Telegraph Speed," *Bell System Technical Journal*, April 1924, p. 324; "Certain Topics in Telegraph Transmission Theory," *A. I. E. E. Trans.*, v. 47, April 1928, p. 617.

[2] Hartley, R. V. L., "Transmission of Information," *Bell System Technical Journal*, July 1928, p. 535.



Claude Shannon

Surprise

You observe draws *x* from a probability distribution *p(x).* How surprised are you at a given outcome? To quantify this, we define a measure *h(p(x))*, called surprise, that satisfies two properties.

Property 1: The surprise of an observation is a decreasing function of the probability of that observation (i.e., unlikely observations are more surprising).

Property 2: The surprise of two independent observations is the sum of the surprises of the individual observations:

$$h(p(x)p(y)) = h(p(x)) + h(p(y))$$

Surprise

You observe draws *x* from a probability distribution *p(x).* How surprised are you at a given outcome? To quantify this, we define a measure *h(p(x))*, called surprise, that satisfies two properties.

Property 1: The surprise of an observation is a decreasing function of the probability of that observation (i.e., unlikely observations are more surprising).

Property 2: The surprise of two independent observations is the sum of the surprises of the individual observations:

$$h(p(x)p(y)) = h(p(x)) + h(p(y))$$

The unique* function satisfying these two properties turns out to be:

$$h(p(x)) = -\log p(x)$$

*unique up to a constant factor, i.e. a change of base of the logarithm

Entropy

- The *entropy* of the distribution *p(x)* is the *expected surprise:*

$$H = \langle h(p(x)) \rangle = -\sum_x p(x) \log p(x)$$

- In other words, entropy quantifies how surprising observations are *on average*.

- Note 1: Surprise pertains to individual observations, entropy to the whole distribution.

- Note 2: The sum implies a discrete distribution – it can be replaced with an integral for continuous distributions, but there are some subtleties involved

Entropy: Example

Consider the binomial distribution. It has two outcomes, *x+* and x-, with $p(x_-) = 1 - p(x_+)$

The entropy is: $H = -(1 - p(x_+)) \log(1 - p(x_+)) - p(x_+) \log p(x_+)$

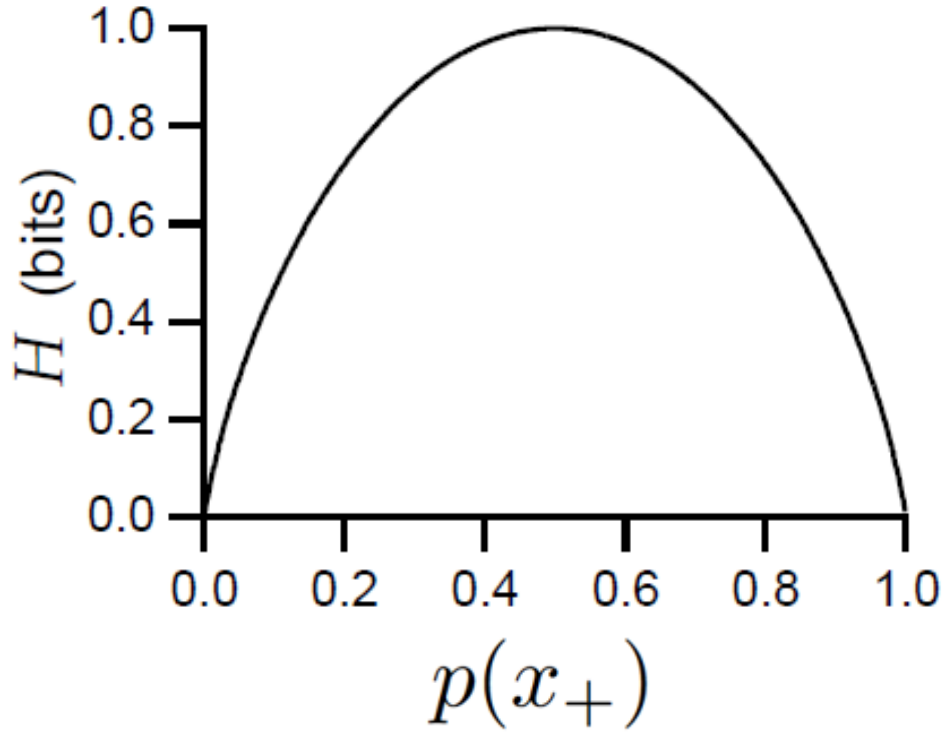
Entropy: Example

Consider the binomial distribution. It has two outcomes, *x+* and x-, with $p(x_-) = 1 - p(x_+)$

The entropy is: $H = -(1 - p(x_+)) \log(1 - p(x_+)) - p(x_+) \log p(x_+)$

e.g., a biased coin flip - entropy is low when the coin lands heads every time, and high when heads/tails are 50/50.

Remember, entropy is *average surprise* – in the biased case, one outcome may be very surprising, but the average surprise is lower



Mutual Information

How much information does one variable convey about another? For example, how much information does a neural response convey about a stimulus?

Mutual information quantifies how much of the variation in the response distribution is explained by variation in the stimulus distribution.

Mutual Information

How much information does one variable convey about another? For example, how much information does a neural response convey about a stimulus?

Mutual information quantifies how much of the variation in the response distribution is explained by variation in the stimulus distribution.

Definition: given two random variables *x* and *y* the mutual information *I(x;y)* is*:*

$$I(x;y) = \underbrace{H(p(y))}_{\text{total entropy}} - \underbrace{\langle H(p(y|x)) \rangle}_{\text{conditional entropy}}$$

The total entropy quantifies the variation in *y,* the conditional entropy quantifies the variation in *y* for fixed *x.* The difference is the variation in *y* that is coupled to variation in *x*.

Mutual Information

Mutual information can be rewritten multiple ways:

$$I(x;y) = \underbrace{H(p(y))}_{\text{total entropy}} - \underbrace{\langle H(p(y|x))\rangle}_{\text{conditional entropy}}$$

$$= -\sum_y p(y)\log p(y) + \langle \sum_y p(y|x)\log p(y|x)\rangle$$

Mutual Information

Mutual information can be rewritten multiple ways:

$$
\begin{aligned}
I(x;y) &= \underbrace{H(p(y))}_{\text{total entropy}} - \underbrace{\langle H(p(y|x))\rangle}_{\text{conditional entropy}} \\
&= -\sum_y p(y)\log p(y) + \langle \sum_y p(y|x)\log p(y|x)\rangle \\
&= -\sum_y p(y)\log p(y) + \sum_{x,y} p(x)p(y|x)\log p(y|x) \\
&= \sum_{x,y} p(x)p(y|x)\log \frac{p(y|x)}{p(y)} \\
&= \sum_{x,y} p(x,y)\log \frac{p(x,y)}{p(x)p(y)} = I(y;x)
\end{aligned}
$$

The last line shows that *I(y;x) = I(x;y)* – mutual information is symmetric.

Mutual Information: Limiting Cases

- If *x* and *y* are independent, mutual information is zero:

$$p(y|x) = p(y) \implies I(x;y) = 0$$

Mutual Information: Limiting Cases

- If *x* and *y* are independent, mutual information is zero:

$$p(y|x) = p(y) \implies I(x;y) = 0$$

- If *y* is perfectly predictable given *x*, i.e. there is deterministic one-to-one mapping, then the mutual information is equal to the entropy of the stimulus distribution:

$$p(y|x) = \begin{cases} 1 & y = y_x \\ 0 & y \neq y_x \end{cases} \implies I(x;y) = H(p(x)) = H(p(y))$$

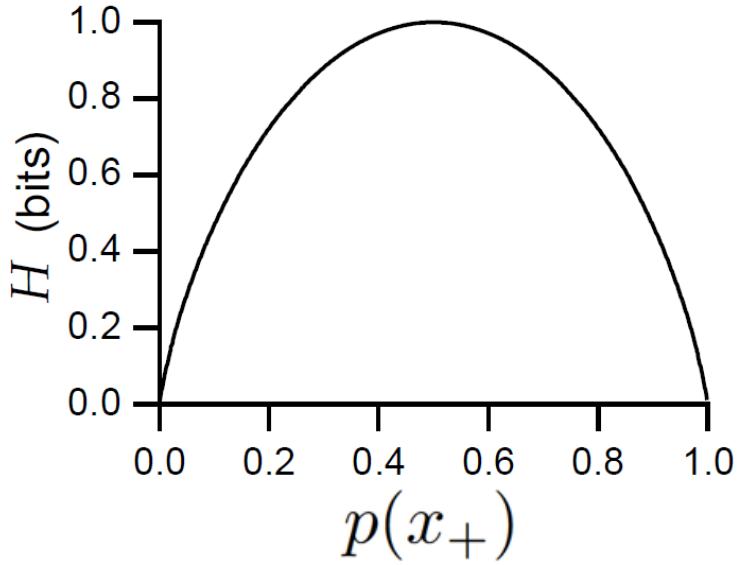
- These examples align with intuition – independent variables do not convey information about one another, whereas perfectly correlated events convey complete information about one another.

Mutual Information: Example

- Let *x* and *y* each follow a binomial distribution, with a noisy mapping from *x* to *y*:

$$p(y_-|x_+) = p_{error}$$

$$p(y_+|x_+) = 1 - p_{error}$$

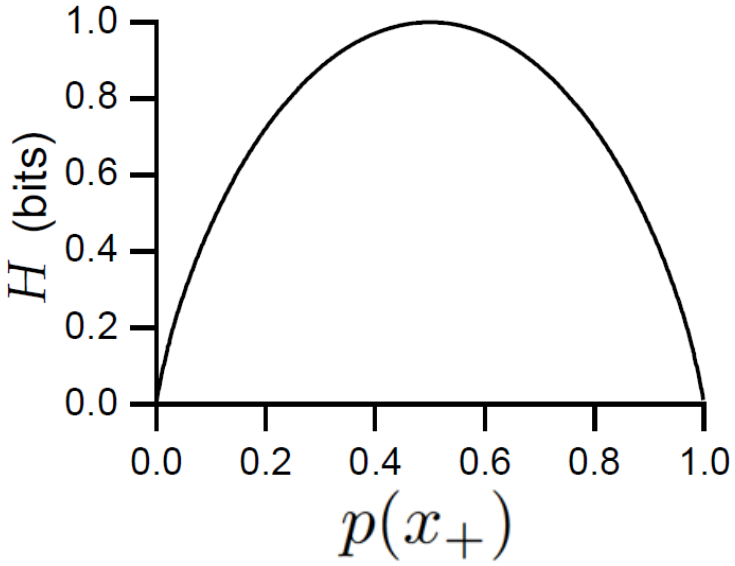


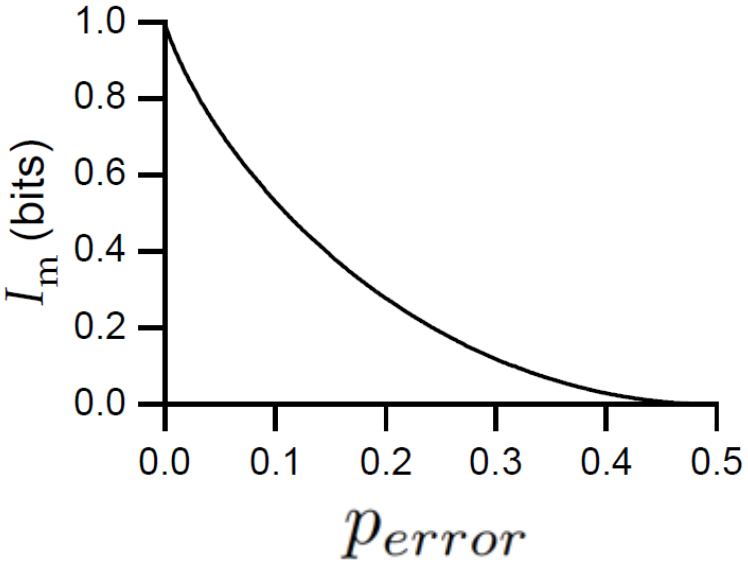
Mutual Information: Example

- Let *x* and *y* each follow a binomial distribution, with a noisy mapping from *x* to *y*:

$$p(y_-|x_+) = p_{error}$$

$$p(y_+|x_+) = 1 - p_{error}$$





- Assume *p(x+) = p(x-) = 0.5*. Then the mutual information is:

$$I(x;y) = 1 + (1 - p_{error}) \log(1 - p_{error}) + p_{error} \log p_{error}$$

Mutual Information: Properties

- Mutual information is symmetric: $I(x;y) = I(y;x)$

Mutual Information: Properties

- Mutual information is symmetric: *I(x;y) = I(y;x)*
- KL-divergence between two distributions *p(x), q(x)* is: $D_{KL}(p,q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$

Mutual Information: Properties

- Mutual information is symmetric: *I(x;y) = I(y;x)*
- KL-divergence between two distributions *p(x), q(x)* is: $D_{KL}(p,q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$

Mutual information is equal to the KL-divergence between the joint and factorised distributions:

$$I(x;y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = D_{KL}(p(x,y); p(x)p(y))$$

Mutual Information: Properties

- Mutual information is symmetric: *I(x;y) = I(y;x)*
- KL-divergence between two distributions *p(x), q(x)* is: $D_{KL}(p,q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$

Mutual information is equal to the KL-divergence between the joint and factorised distributions:

$$I(x;y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = D_{KL}(p(x,y); p(x)p(y))$$

- Corollaries based on properties of KL: mutual information is 1) non-negative 2) zero only when *x* and *y* are independent 3) a measure of distance from independence.

The Data Processing Inequality

Statement of theorem: Let $x \rightarrow y \rightarrow z$ be a set of 3 random variables, with arbitrary (random or deterministic) mappings represented by arrows. Then $I(x;z) \leq I(x;y)$

The Data Processing Inequality

Statement of theorem: Let $x \rightarrow y \rightarrow z$ be a set of 3 random variables, with arbitrary (random or deterministic) mappings represented by arrows. Then $I(x;z) \leq I(x;y)$

For example, consider that *x* is a visual stimulus, *y* is the response of the retina, and *z* is the joint response of all neurons in the brain. Then this theorem proves that there is more information about the visual stimulus in the retina than there is in the brain.

The Data Processing Inequality

Statement of theorem: Let $x \rightarrow y \rightarrow z$ be a set of 3 random variables, with arbitrary (random or deterministic) mappings represented by arrows. Then $I(x;z) \leq I(x;y)$

For example, consider that *x* is a visual stimulus, *y* is the response of the retina, and *z* is the joint response of all neurons in the brain. Then this theorem proves that there is more information about the visual stimulus in the retina than there is in the brain.

Neural processing can reformat representations, combine different streams of information, filter out noise, etc., but information can never increase!

Important assumptions:

1) *x* influences *y*, *y* influences *z*, but no feedback loops (Markov chain)

2) Ignores time. But the theorem still applies if we define {x,y,z} as the entire history of {x(t),y(t),z(t)}

Summary of Information Theory

- Information theory quantifies communication of signals through a noisy channel
- Three important quantities: surprise, entropy, and mutual information
- Mutual information quantifies how accurately a stimulus can be reconstructed from a neural response (or reduction of uncertainty about stimulus open observing the response)
- The data processing inequality tells us what the brain *can't* do (i.e., increase information)
- What *can* the brain do? Communicate efficiently under capacity/resource constraints!

The Efficient Coding Hypothesis

Information theory provides a normative framework for understanding sensory systems.

Why do retinal ganglion cells have ON-OFF receptive fields? Why do they adapt in the way they do to light vs dark? Evolution must have chosen something useful.

The efficient coding theory postulates that these properties are optimal, given the natural statistics of sensory input and the constraints the brain works under (e.g., physiological noise, energetic costs).

It turns out we can derive, from first principles, something that looks roughly like what we find in the nervous system (works best near sensory periphery).

13

H. B. BARLOW
Physiological Laboratory, Cambridge University

Possible Principles Underlying the Transformations of Sensory Messages

A wing would be a most mystifying structure if one did not know that birds flew. One might observe that it could be extended a considerable distance, that it had a smooth covering of feathers with conspicuous markings, that it was operated by powerful muscles, and that strength and lightness were prominent features of its construction. These are important facts, but by themselves they do not tell us that birds fly. Yet without knowing this, and without understanding something of the principles of flight, a more detailed examination of the wing itself would probably be unrewarding. I think that we may be at an analogous point in our understanding of the sensory side of the central nervous system. We have got our first batch of facts from the anatomical, neurophysiological, and psychophysical study of sensation and perception, and now we need ideas about what operations are performed by the various structures we have examined. For the bird's wing we can say that it accelerates downwards the air flowing past it and so derives an upward force which supports the weight of the bird; what would be a similar summary of the most important operation performed at a sensory relay?

Maximisation of Mutual Information

Suppose we wish to maximise the mutual information between a set of stimuli *s* and neural responses *r*:

$$I(s;r) = \underbrace{H(p(r))}_{\text{response entropy}} - \underbrace{\langle H(p(r|s)) \rangle}_{\text{noise entropy}}$$

This involves a trade-off between two terms - minimising noise entropy and maximising response entropy.

This could be solved trivially – e.g. why not simply set *r=s*?

Answer: there are resource constraints, noise, bottlenecks, etc.

Histogram Equalisation

Consider encoding of a stimulus *s* by a single neuron with firing rate *r*. To make life simple, we assume that noise entropy is small, so that we need only maximise response entropy:

$$I(s;r) \approx H(p(r)) = -\int p(r) \log p(r) dr$$

Histogram Equalisation

Consider encoding of a stimulus *s* by a single neuron with firing rate *r.* To make life simple, we assume that noise entropy is small, so that we need only maximise response entropy:

$$I(s;r) \approx H(p(r)) = -\int p(r) \log p(r) dr$$

But we assume that the neuron can only respond with rates above 0 and below *rmax.* This gives the constraint:

$$\int_0^{r_{max}} p(r) dr = 1$$

Histogram Equalisation

Consider encoding of a stimulus *s* by a single neuron with firing rate *r.* To make life simple, we assume that noise entropy is small, so that we need only maximise response entropy:

$$I(s;r) \approx H(p(r)) = -\int p(r) \log p(r) dr$$

But we assume that the neuron can only respond with rates above 0 and below *rmax.* This gives the constraint:

$$\int_0^{r_{max}} p(r) dr = 1$$

This is a constrained optimisation problem, and can be solved using Lagrange multipliers (see Dayan and Abbott, Ch 4). The solution is:

$$p(r) = \frac{1}{r_{max}} \qquad H(p(r)) = \log r_{max}$$

Histogram Equalisation

This solution sets all firing rates as equally likely - this is a well-known signal processing technique called histogram equalisation.

What does that tell us about the stimulus encoding? How does *r* relate to *s*?

Histogram Equalisation

This solution sets all firing rates as equally likely - this is a well-known signal processing technique called histogram equalisation.

What does that tell us about the stimulus encoding? How does *r* relate to *s*?

Suppose we have a stimulus distribution *p(s)* encoded as *r=f(s)* (remember: we have assumed noise-free encoding). Given that *p(r)=1/rmax*, we have (using the rule for *change of random variables*):

$$p_s(s) = p_r(r = f(s))|f'(s)|$$

Note: we write subscript *r* and *s* to clarify that these are different distributions. We assume that *f* is monotonic increasing.

Histogram Equalisation

This solution sets all firing rates as equally likely - this is a well-known signal processing technique called histogram equalisation.

What does that tell us about the stimulus encoding? How does *r* relate to *s*?

Suppose we have a stimulus distribution *p(s)* encoded as *r=f(s)* (remember: we have assumed noise-free encoding). Given that *p(r)=1/rmax*, we have (using the rule for *change of random variables*):

$$p_s(s) = p_r(r = f(s))|f'(s)| = \frac{f'(s)}{r_{max}}$$

Note: we write subscript *r* and *s* to clarify that these are different distributions. We assume that *f* is monotonic increasing.

Histogram Equalisation

This solution sets all firing rates as equally likely - this is a well-known signal processing technique called histogram equalisation.

What does that tell us about the stimulus encoding? How does *r* relate to *s*?

Suppose we have a stimulus distribution *p(s)* encoded as *r=f(s)* (remember: we have assumed noise-free encoding). Given that *p(r)=1/rmax*, we have (using the rule for *change of random variables*):

$$p_s(s) = p_r(r = f(s))|f'(s)| = \frac{f'(s)}{r_{max}} \implies f(s) = r_{max} \int_{s_{min}}^{s} p(s')ds'$$

The tuning curve of the neuron is the cumulative distribution of the stimulus!

Note: we write subscript *r* and *s* to clarify that these are different distributions. We assume that *f* is monotonic increasing.

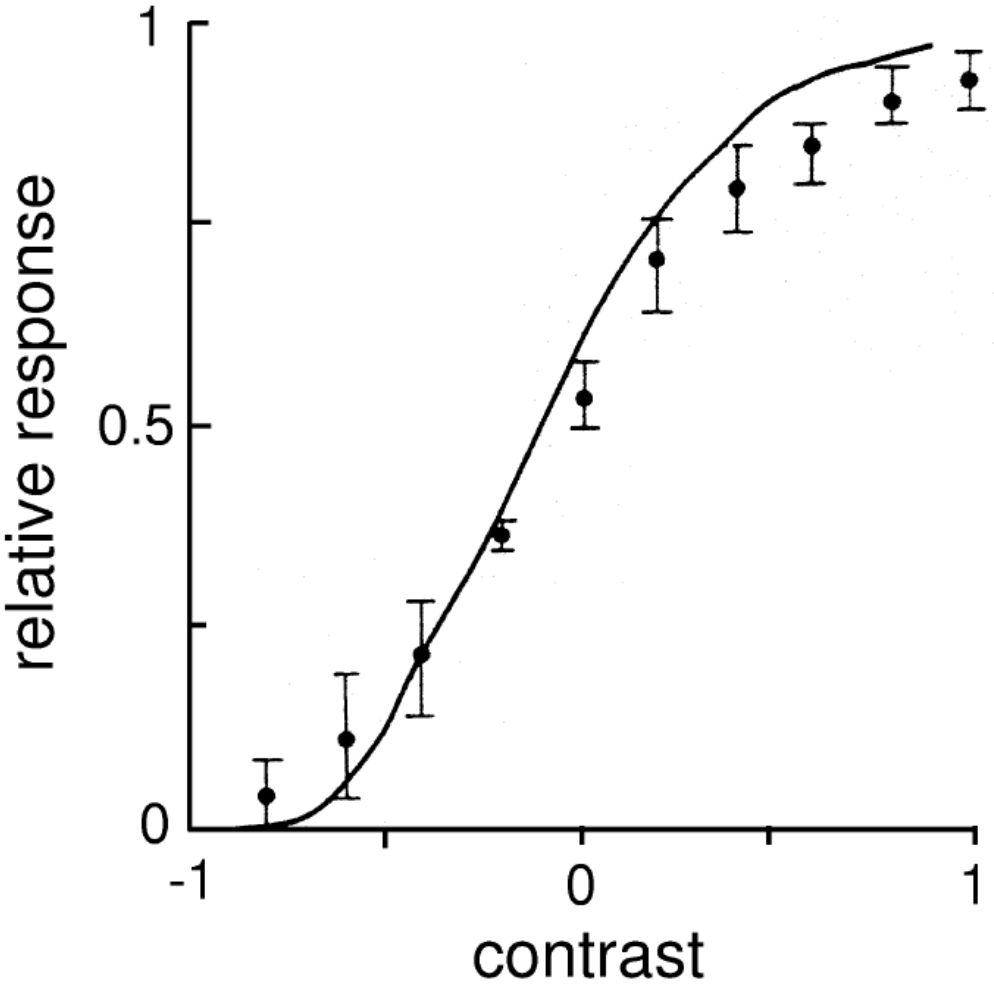
Histogram Equalisation in the Fly Visual System

One can test for histogram equalisation by:

1. Measuring the natural sensory statistics in an organism's environment

2. Computing the cumulative probability distribution of the stimulus *s*

3. Comparing this cumulative distribution to the tuning curves measured in the organism



The figures shown such a comparison in the fly visual system (error bars are neural data, solid curve is cumulative distribution of stimulus contrast).

Laughlin 1981

Summary: Histogram Equalisation

- We have considered a simple version of efficient coding: noise-free, single neuron, constraint on range of firing rates
- The solution is to set all firing rates in range equally likely
- This predicts that the tuning curve is the cumulative distribution of the stimulus
- We found evidence for such an encoding in the fly visual system
- But we’re missing: noise, multiple neurons, various other constraints

Extension to Populations of Neurons: Independent Coding

What if we have multiple neurons? Response entropy for a population of neurons is:

$$H(p(\mathbf{r})) = -\int p(\mathbf{r}) \log p(\mathbf{r}) d\mathbf{r}$$

Extension to Populations of Neurons: Independent Coding

What if we have multiple neurons? Response entropy for a population of neurons is:

$$H(p(\mathbf{r})) = -\int p(\mathbf{r}) \log p(\mathbf{r}) d\mathbf{r}$$

The total response entropy is always less than that of a statistically independent population:

$$H(p(\mathbf{r})) \leq \sum_i H(p(r_i))$$

Extension to Populations of Neurons: Independent Coding

What if we have multiple neurons? Response entropy for a population of neurons is:

$$H(p(\mathbf{r})) = -\int p(\mathbf{r}) \log p(\mathbf{r}) d\mathbf{r}$$

The total response entropy is always less than that of a statistically independent population:

$$H(p(\mathbf{r})) \leq \sum_i H(p(r_i))$$

$$H(p(\mathbf{r})) = \sum_i H(p(r_i)) \iff p(\mathbf{r}) = \prod_i p(r_i)$$

Thus, we can maximise entropy by finding a code where neurons are statistically independent. This is not easy in general (e.g., for natural images).

Application: Decorrelation by Retinal Ganglion Cells

- Why do retinal ganglion cells have centre-surround receptive fields?
- Pixels in a natural image are correlated. Thus, one pixel can be predicted from others. This is a redundant, and therefore an inefficient code.
- What if we try to find a code where the image is represented by independent features?
- For example, can we find a filter that decorrelates (“whitens”) natural images? If so, does that filter resemble retinal ganglion cell receptive fields?
- Note: maximising entropy directly is too hard here, but decorrelation maximises entropy for Gaussian distributions, and typically increases entropy for non-Gaussian ones…

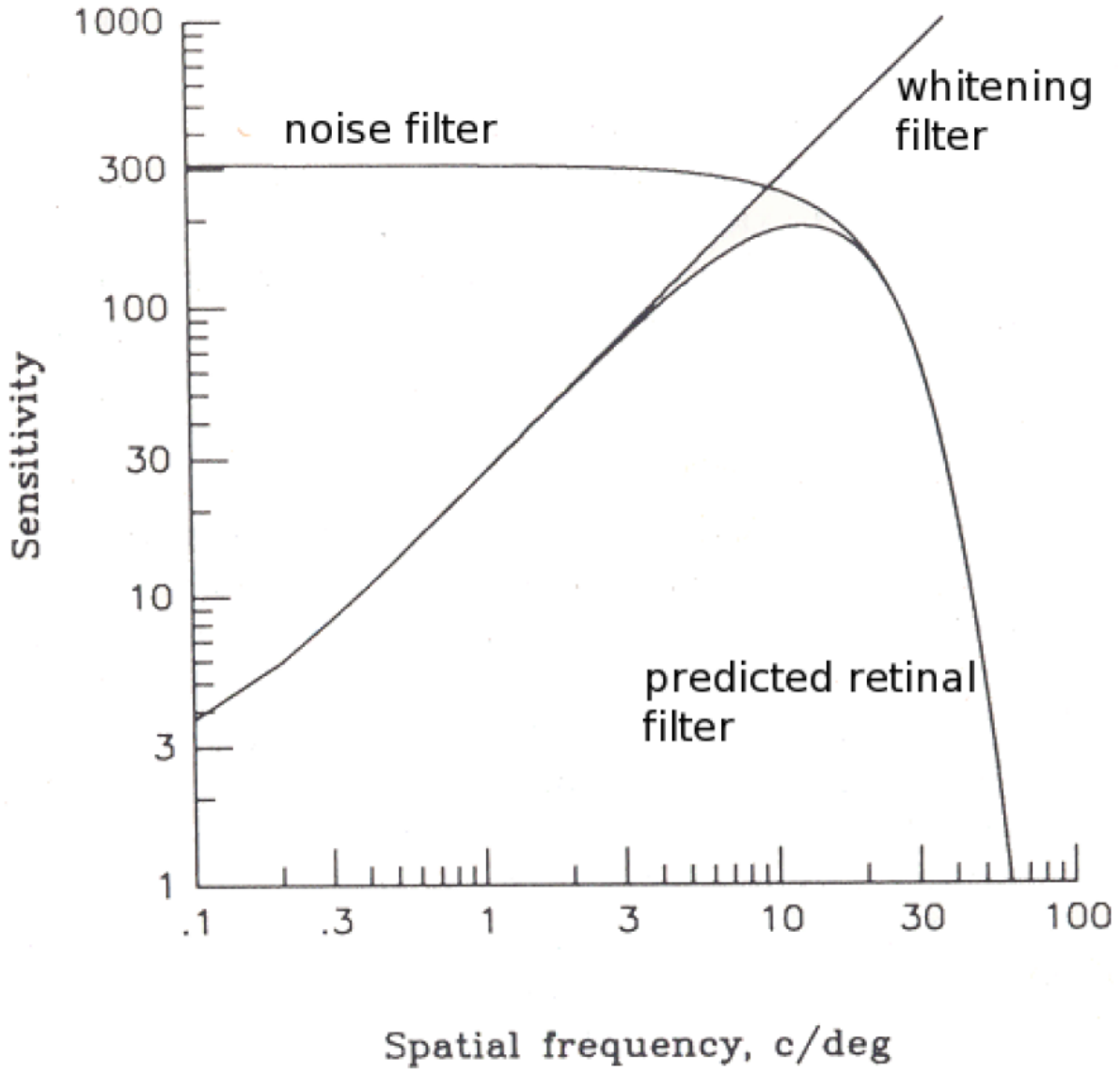
Decorrelation by Retinal Ganglion Cells

We can compute the whitening filter from the statistics of natural images
(see Dayan and Abbott Ch 4 for derivation)

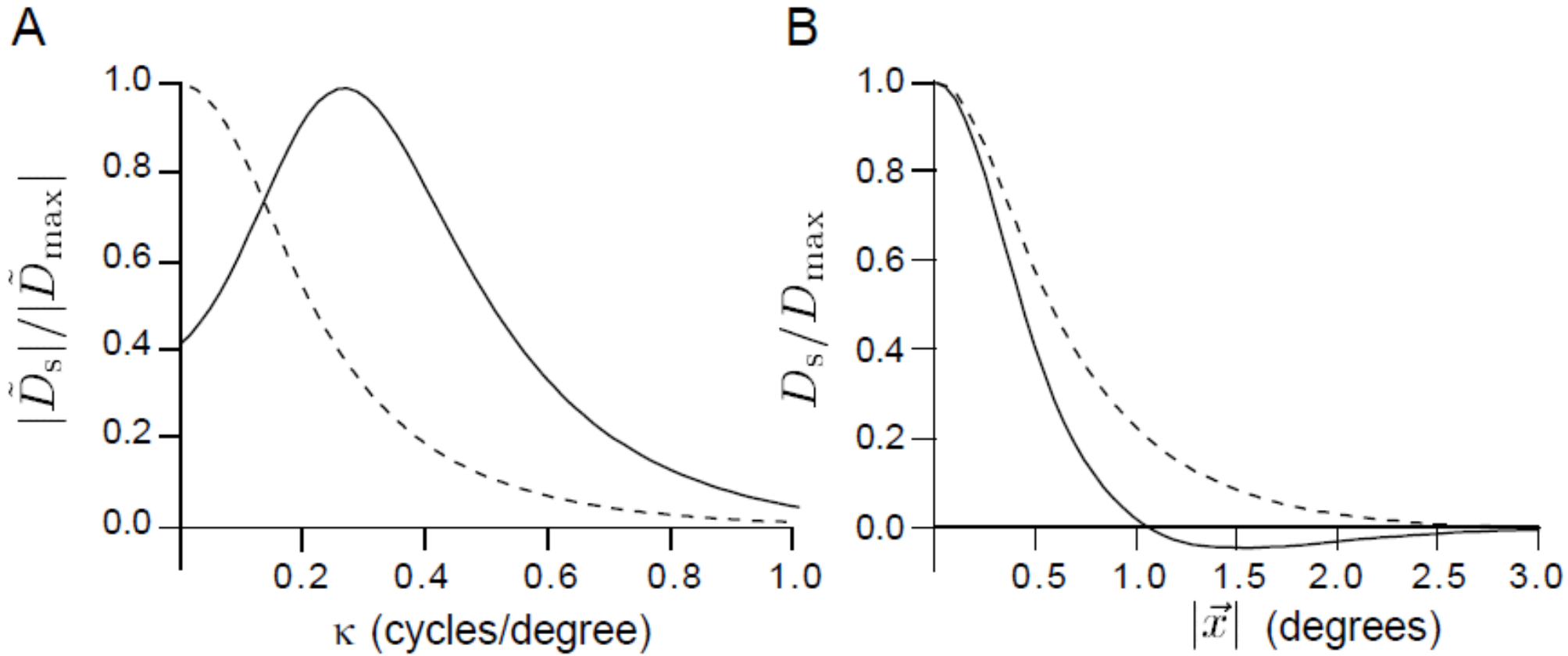
The whitening filter grows exponentially at high frequencies.

But we have assumed a noise-free encoding – in reality, there is noise at high frequencies (photon and retinal noise).

The optimal filter trades off whitening against noise removal and falls off at high frequencies.



Decorrelation by Retinal Ganglion Cells



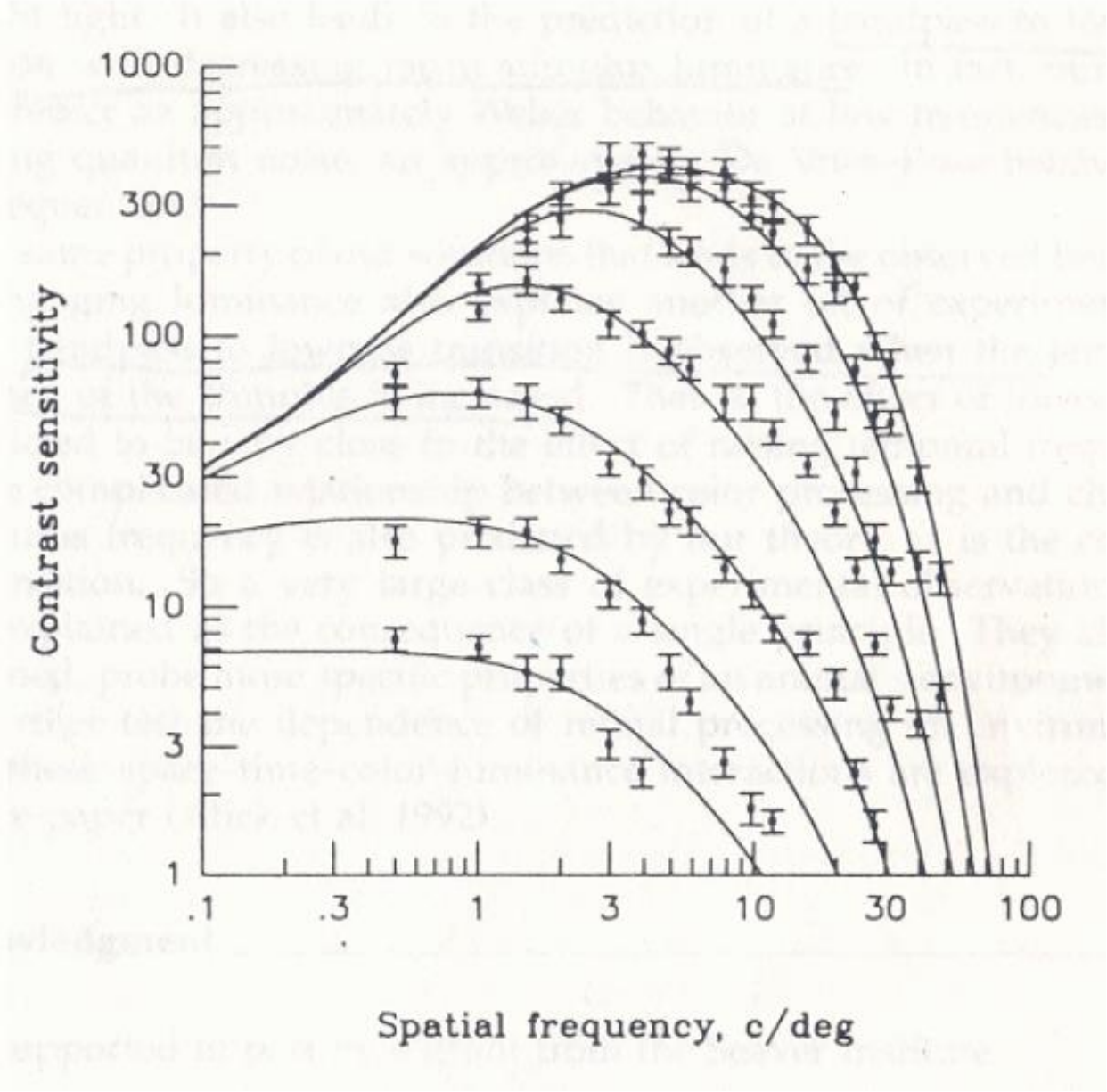
A: The optimal filter is bandpass at low noise (solid) and low pass at high noise (dashed).

B: The optimal filter looks like a difference of Gaussians at low noise and a Gaussian at high noise.

Decorrelation by Retinal Ganglion Cells

RGC receptive fields adapt to different lighting conditions.

Optimal filters at different noise levels can explain this adaptation
(low light = high noise)



Summary: Whitening by Retinal Ganglion Cells

- Retinal ganglion cells have difference of Gaussian receptive fields that act as bandpass filters.
- We asked whether these filters can be derived from natural image statistics, assuming they remove correlations (whiten) to encode more efficiently.
- The whitening filter predicts a high-pass filter
- When high-frequency noise is assumed, the optimal filter is bandpass and resembles those of RGCs, including adaptation to lighting conditions.

Limitations of Information-Theoretic Approaches

Requires lots of data, not possible in practice due to experimental limitations, therefore have to use approximations and strong assumptions.

Does not address the meaning/purpose/computation. Only quantifies correspondences between stimuli and neural responses. E.g., what about object recognition?

Does not tell us whether/how/for what information is actually *used* by the brain.

Views the brain as encoding and then decoding a stimulus – why would the brain do that? Might be reasonable in e.g., the retina, but less clear in later stages of processing.

Summary of Lecture

Neural coding is the relationship between states in the world and in the brain

Information theory can be used to quantify the amount of information in a neural code, and to find optimal codes under constraints

The efficient coding hypothesis postulates that the nervous system employs an optimal code, given the natural stimulus statistics and resource constraints

Success stories include: Histogram equalisation in fly visual system, decorrelation in retina

Bibliography

Neural Computation lecture notes chapter 7-9. [the visual system, coding, information theory]

Dayan and Abbott Chapter 4. [information theory in the neuroscience context]

Cover and Thomas (Elements of Information Theory) [general information theory]

Rieke (Spikes: Exploring the Neural Code) [information in spike trains]