

## 4.2 Entropy lost and information gained

Section last updated  
April 20, 2009.

Returning to the conversation between Max and Allan, we assumed that Max would receive a complete answer to his question, and hence that all his uncertainty would be removed. This is an idealization, of course. The more natural description is that, for example, the world can take on many states  $W$ , and by observing data  $D$  we learn something but not everything about  $W$ . Before we make our observations, we know only that states of the world are chosen from some distribution  $P(W)$ , and this distribution has an entropy  $S(W)$ . Once we observe some particular datum  $D$ , our (hopefully improved) knowledge of  $W$  is described by the conditional distribution  $P(W|D)$ , and this has an entropy  $S(W|D)$  that is smaller than  $S(W)$  if we have reduced our uncertainty about the state of the world by virtue of our observations. We identify this reduction in entropy as the information that we have gained about  $W$ .

Perhaps this is the point to note that a single observation  $D$  is not, in fact, guaranteed to provide positive information [see, for example, DeWeese and Meister 1999]. Consider, for instance, data which tell us that all of our previous measurements have larger error bars than we thought: clearly such data, at an intuitive level, reduce our knowledge about the world and should be associated with a negative information. Another way to say this is that some data points  $D$  will increase our uncertainty about state  $W$  of the world, and hence for these particular data the conditional distribution  $P(W|D)$  has a larger entropy than the prior distribution  $P(W)$ . If we identify information with the reduction in entropy,  $I_D = S(W) - S(W|D)$ , then such data points are associated unambiguously with negative information. On the other hand, we might hope that, on average, gathering data corresponds to gaining information: although single data points can increase our uncertainty, the average over all data points does not.

If we average over all possible data—weighted, of course, by their probability of occurrence  $P(D)$ —we obtain the average information that  $D$  provides about  $W$ :

$$I(D \rightarrow W) = S(W) - \sum_D P(D) S(W|D) \quad (4.29)$$

$$\begin{aligned} &= - \sum_W P(W) \log_2 P(W) \\ &\quad - \sum_D P(D) \left[ - \sum_W P(W|D) \log_2 P(W|D) \right] \quad (4.30) \end{aligned}$$

$$\begin{aligned}
&= -\sum_W \sum_D P(W, D) \log_2 P(W) \\
&\quad + \sum_W \sum_D P(W|D)P(D) \log_2 P(W|D) \quad (4.31)
\end{aligned}$$

$$\begin{aligned}
&= -\sum_W \sum_D P(W, D) \log_2 P(W) \\
&\quad + \sum_W \sum_D P(W, D) \log_2 P(W|D) \quad (4.32)
\end{aligned}$$

$$= \sum_W \sum_D P(W, D) \log_2 \left[ \frac{P(W|D)}{P(W)} \right] \quad (4.33)$$

$$= \sum_W \sum_D P(W, D) \log_2 \left[ \frac{P(W, D)}{P(W)P(D)} \right]. \quad (4.34)$$

We see that, after all the dust settles, the information which  $D$  provides about  $W$  is symmetric in  $D$  and  $W$ . This means that we can also view the state of the world as providing information about the data we will observe, and this information is, on average, the same as the data will provide about the state of the world. This ‘information provided’ is therefore often called the mutual information, and this symmetry will be very important in subsequent discussions; to remind ourselves of this symmetry we write  $I(D; W)$  rather than  $I(D \rightarrow W)$ .

One consequence of the symmetry or mutuality of information is that we can write

$$I(D; W) = S(W) - \sum_D P(D)S(W|D) \quad (4.35)$$

$$= S(D) - \sum_W P(W)S(D|W). \quad (4.36)$$

If we consider only discrete sets of possibilities then entropies are positive (or zero), so that these equations imply

$$I(D; W) \leq S(W) \quad (4.37)$$

$$I(D; W) \leq S(D). \quad (4.38)$$

The first equation tells us that by observing  $D$  we cannot learn more about the world than there is entropy in the world itself. This makes sense: entropy measures the number of possible states that the world can be in, and we cannot learn more than we would learn by reducing this set of possibilities down to one unique state. Although sensible (and, of course, true), this is

not a terribly powerful statement: seldom are we in the position that our ability to gain knowledge is limited by the lack of possibilities in the world around us, although there is a tradition of studying the nervous system as it responds to highly simplified signals, and under these conditions the lack of possibilities in the world can be a significant limitation, substantially confounding the interpretation of experiments.

Equation (4.38), however, is much more powerful. It says that, whatever may be happening in the world, we can never learn more than the entropy of the distribution that characterizes our data. Thus, if we ask how much we can learn about the world by taking readings from a wind detector on top of the roof, we can place a bound on the amount we learn just by taking a very long stream of data, using these data to estimate the distribution  $P(D)$ , and then computing the entropy of this distribution.

The entropy of our observations thus limits how much we can learn no matter what question we were hoping to answer, and so we can think of the entropy as setting (in a slight abuse of terminology) the capacity of the data  $D$  to provide or to convey information. As an example, the entropy of neural responses sets a limit to how much information a neuron can provide about the world, and we can estimate this limit even if we don't yet understand what it is that the neuron is telling us (or the rest of the brain).

---

**Problem 42: Maximally informative experiments.** Imagine that we are trying to gain information about the correct theory  $T$  describing some set of phenomena. At some point, our relative confidence in one particular theory is very high; that is,  $P(T = T_*) > F \cdot P(T \neq T_*)$  for some large  $F$ . On the other hand, there are many possible theories, so our absolute confidence in the theory  $T_*$  might nonetheless be quite low,  $P(T = T_*) \ll 1$ . Suppose we follow the 'scientific method' and design an experiment that has a yes or no answer, and this answer is perfectly correlated with the correctness of theory  $T_*$ , but uncorrelated with the correctness of any other possible theory—our experiment is designed specifically to test or falsify the currently most likely theory. What can you say about how much information you expect to gain from such a measurement? Suppose instead that you are completely irrational and design an experiment that is irrelevant to testing  $T_*$  but has the potential to eliminate many (perhaps half) of the alternatives. Which experiment is expected to be more informative? Although this is a gross cartoon of the scientific process, it is not such a terrible model of a game like "twenty questions." It is interesting to ask whether people play such question games following strategies that might seem irrational but nonetheless serve to maximize information gain. Related but distinct criteria for optimal experimental design have been developed in the statistical literature.

---

We now want to look at information transmission in the presence of noise, connecting back a bit to what we discussed in earlier parts of the of course. Imagine that we are interested in some signal  $x$ , and we have a detector that generates data  $y$  which is linearly related to the signal but corrupted by added noise:

$$y = gx + \xi. \quad (4.39)$$

It seems reasonable in many systems to assume that the noise arises from many added sources (e.g., the Brownian motion of electrons in a circuit) and hence has a Gaussian distribution because of the central limit theorem. We will also start with the assumption that  $x$  is drawn from a Gaussian distribution just because this is a simple place to start; we will see that we can use the maximum entropy property of Gaussians to make some more general statements based on this simple example. The question, then, is how much information observations on  $y$  provide about the signal  $x$ .

Let us formalize our assumptions. The statement that  $\xi$  is Gaussian noise means that once we know  $x$ ,  $y$  is Gaussian distributed around a mean value of  $gx$ :

$$P(y|x) = \frac{1}{\sqrt{2\pi\langle\xi^2\rangle}} \exp\left[-\frac{1}{2\langle\xi^2\rangle}(y-gx)^2\right]. \quad (4.40)$$

Our simplification is that the signal  $x$  also is drawn from a Gaussian distribution,

$$P(x) = \frac{1}{\sqrt{2\pi\langle x^2\rangle}} \exp\left[-\frac{1}{2\langle x^2\rangle}x^2\right], \quad (4.41)$$

and hence  $y$  itself is Gaussian,

$$P(y) = \frac{1}{\sqrt{2\pi\langle y^2\rangle}} \exp\left[-\frac{1}{2\langle y^2\rangle}y^2\right] \quad (4.42)$$

$$\langle y^2 \rangle = g^2\langle x^2 \rangle + \langle \xi^2 \rangle. \quad (4.43)$$

To compute the information that  $y$  provides about  $x$  we use Eq. (4.34):

$$I(y \rightarrow x) = \int dy \int dx P(x, y) \log_2 \left[ \frac{P(x, y)}{P(x)P(y)} \right] \text{ bits} \quad (4.44)$$

$$= \frac{1}{\ln 2} \int dy \int dx P(x, y) \ln \left[ \frac{P(y|x)}{P(y)} \right] \quad (4.45)$$

$$= \frac{1}{\ln 2} \left\langle \ln \left[ \frac{\sqrt{2\pi\langle y^2 \rangle}}{\sqrt{2\pi\langle \xi^2 \rangle}} \right] - \frac{1}{2\langle \xi^2 \rangle}(y-gx)^2 + \frac{1}{2\langle y^2 \rangle}y^2 \right\rangle, \quad (4.46)$$

where by  $\langle \dots \rangle$  we understand an expectation value over the joint distribution  $P(x, y)$ . Now in Eq. (4.46) we can see that the first term is the expectation value of a constant, which is just the constant. The third term involves the expectation value of  $y^2$  divided by  $\langle y^2 \rangle$ , so we can cancel numerator and denominator. In the second term, we can take the expectation value first of  $y$  with  $x$  fixed, and then average over  $x$ , but since  $y = gx + \xi$  the numerator is just the mean square fluctuation of  $y$  around its mean value, which again cancels with the  $\langle \xi^2 \rangle$  in the denominator. So we have, putting the three terms together,

$$I(y \rightarrow x) = \frac{1}{\ln 2} \left[ \ln \sqrt{\frac{\langle y^2 \rangle}{\langle \xi^2 \rangle}} - \frac{1}{2} + \frac{1}{2} \right] \quad (4.47)$$

$$= \frac{1}{2} \log_2 \left( \frac{\langle y^2 \rangle}{\langle \xi^2 \rangle} \right) \quad (4.48)$$

$$= \frac{1}{2} \log_2 \left( 1 + \frac{g^2 \langle x^2 \rangle}{\langle \xi^2 \rangle} \right) \text{ bits.} \quad (4.49)$$

Although it may seem like useless algebra, I would like to rewrite this result a little bit. Rather than thinking of our detector as adding noise after generating the signal  $gx$ , we can think of it as adding noise directly to the input, and then transducing this corrupted input:

$$y = g(x + \eta_{\text{eff}}), \quad (4.50)$$

where, obviously,  $\eta_{\text{eff}} = \xi/g$ . Note that the “effective noise”  $\eta_{\text{eff}}$  is in the same units as the input  $x$ ; this is called ‘referring the noise to the input’ and is a standard way of characterizing detectors, amplifiers and other devices.<sup>5</sup> Written in terms of the effective noise level, the information transmission takes a simple form,

$$I(y \rightarrow x) = \frac{1}{2} \log_2 \left( 1 + \frac{\langle x^2 \rangle}{\langle \eta_{\text{eff}}^2 \rangle} \right) \text{ bits,} \quad (4.51)$$

or

$$I(y \rightarrow x) = \frac{1}{2} \log_2(1 + SNR), \quad (4.52)$$

---

<sup>5</sup>As an example, if we build a photodetector it is not so useful to quote the noise level in Volts at the output—we want to know how this noise limits our ability to detect dim lights. Similarly, when we characterize a neuron that uses a stream of pulses to encode a continuous signal, we don’t really want to know the variance in the pulse rate (although this is widely discussed); we want to know how noise in the neural response limits precision in estimating the real signal, and this amounts to defining an effective noise level in the units of the signal itself. In the present case this is just a matter of dividing, but generally it is a more complex task.

where the signal to noise ratio is the ratio of the variance in the signal to the variance of the effective noise,  $SNR = \langle x^2 \rangle / \langle \eta_{\text{eff}}^2 \rangle$ .

The result in Eq. (4.52) is easy to picture: When we start, the signal is spread over a range  $\delta x_0 \sim \langle x^2 \rangle^{1/2}$ , but by observing the output of our detector we can localize the signal to a small range  $\delta x_1 \sim \langle \eta_{\text{eff}}^2 \rangle^{1/2}$ , and the reduction in entropy is  $\sim \log_2(\delta x_0 / \delta x_1) \sim (1/2) \cdot \log_2(SNR)$ , which is approximately the information gain.

**Problem 43: A small point.** Try to understand why the simple argument in the preceding paragraph, which seems sensible, doesn't give the exact answer for the information gain at small  $SNR$ .

As a next step consider the case where we observe several variables  $y_1, y_2, \dots, y_K$  in the hopes of learning about the same number of underlying signals  $x_1, x_2, \dots, x_K$ . The equations analogous to Eq. (4.39) are then

$$y_i = g_{ij}x_j + \xi_i, \quad (4.53)$$

with the usual convention that we sum over repeated indices. The Gaussian assumptions are that each  $x_i$  and  $\xi_i$  has zero mean, but in general we have to think about arbitrary covariance matrices,

$$S_{ij} = \langle x_i x_j \rangle \quad (4.54)$$

$$N_{ij} = \langle \xi_i \xi_j \rangle. \quad (4.55)$$

The relevant probability distributions are

$$P(\{x_i\}) = \frac{1}{\sqrt{(2\pi)^K \det S}} \exp \left[ -\frac{1}{2} x_i \cdot (S^{-1})_{ij} \cdot x_j \right] \quad (4.56)$$

$$P(\{y_i\}|\{x_i\}) = \frac{1}{\sqrt{(2\pi)^K \det N}} \times \exp \left[ -\frac{1}{2} (y_j - g_{jk}x_k) \cdot (N^{-1})_{ij} \cdot (y_j - g_{jm}x_m) \right], \quad (4.57)$$

where again the summation convention is used;  $\det S$  denotes the determinant of the matrix  $S$ ,  $(S^{-1})_{ij}$  is the  $ij$  element in the inverse of the matrix  $S$ , and similarly for the matrix  $N$ .

To compute the mutual information we proceed as before. First we find  $P(\{y_i\})$  by doing the integrals over the  $x_i$ ,

$$P(\{y_i\}) = \int d^K x P(\{y_i\}|\{x_i\})P(\{x_i\}), \quad (4.58)$$

and then we write the information as an expectation value,

$$I(\{y_i\} \rightarrow \{x_i\}) = \left\langle \log_2 \left[ \frac{P(\{y_i\}|\{x_i\})}{P(\{y_i\})} \right] \right\rangle, \quad (4.59)$$

where  $\langle \dots \rangle$  denotes an average over the joint distribution  $P(\{y_i\}, \{x_i\})$ . As in Eq. (4.46), the logarithm can be broken into several terms such that the expectation value of each one is relatively easy to calculate. Two of three terms cancel, and the one which survives is related to the normalization factors that come in front of the exponentials. After the dust settles we find

$$I(\{y_i\} \rightarrow \{x_i\}) = \frac{1}{2} \text{Tr} \log_2 [\mathbf{1} + N^{-1} \cdot (g \cdot S \cdot g^T)], \quad (4.60)$$

where  $\text{Tr}$  denotes the trace of a matrix,  $\mathbf{1}$  is the unit matrix, and  $g^T$  is the transpose of the matrix  $g$ .

The matrix  $g \cdot S \cdot g^T$  describes the covariance of those components of  $y$  that are contributed by the signal  $x$ . We can always rotate our coordinate system on the space of  $y$ s to make this matrix diagonal, which corresponds to finding the eigenvectors and eigenvalues of the covariance matrix; these eigenvectors are also called “principal components.” For a Gaussian distribution, the eigenvectors describe directions in the space of  $y$  which are fluctuating independently, and the eigenvalues are the variances along each of these directions. If the covariance of the noise is diagonal in the same coordinate system, then the matrix  $N^{-1} \cdot (g \cdot S \cdot g^T)$  is diagonal and the elements along the diagonal are the signal to noise ratios along each independent direction. Taking the  $\text{Tr} \log$  is equivalent to computing the information transmission along each direction using Eq. (4.52), and then summing the results.

An important case is when the different variables  $x_i$  represent a signal sampled at several different points in time. Then there is some underlying continuous function  $x(t)$ , and in place of the discrete Eq. (4.53) we have the continuous linear response of the detector to input signals,

$$y(t) = \int dt' M(t-t')x(t') + \xi(t). \quad (4.61)$$

In this continuous case the analog of the covariance matrix  $\langle x_i x_j \rangle$  is the correlation function  $\langle x(t)x(t') \rangle$ . We are usually interested in signals (and

noise) that are stationary. This means—as discussed previously, and in Appendix B—that all statistical properties of the signal are invariant to translations in time: a particular pattern of wiggles in the function  $x(t)$  is equally likely to occur at any time. Thus, the correlation function which could in principle depend on two times  $t$  and  $t'$  depends only on the time difference,

$$\langle x(t)x(t') \rangle = C_x(t - t'). \quad (4.62)$$

The correlation function generalizes the covariance matrix to continuous time, but we have seen that it can be useful to diagonalize the covariance matrix, thus finding a coordinate system in which fluctuations in the different directions are independent. From previous lectures we know that the answer is to go into a Fourier representation, where (in the Gaussian case) different Fourier components are independent and their variances are (up to normalization) the power spectra.

To complete the analysis of the continuous time Gaussian channel described by Eq. (4.61), we again refer noise to the input by writing

$$y(t) = \int dt' M(t - t')[x(t') + \eta_{\text{eff}}(t')]. \quad (4.63)$$

If both signal and effective noise are stationary, then each has a power spectrum; let us denote the power spectrum of the effective noise  $\eta_{\text{eff}}$  by  $N_{\text{eff}}(\omega)$  and the power spectrum of  $x$  by  $S_x(\omega)$  as usual. There is a signal to noise ratio at each frequency,

$$SNR(\omega) = \frac{S_x(\omega)}{N_{\text{eff}}(\omega)}, \quad (4.64)$$

and since we have diagonalized the problem by Fourier transforming, we can compute the information just by adding the contributions from each frequency component, so that

$$I[y(t) \rightarrow x(t)] = \frac{1}{2} \sum_{\omega} \log_2[1 + SNR(\omega)]. \quad (4.65)$$

Finally, to compute the frequency sum, we recall that

$$\sum_{\mathbf{n}} f(\omega_{\mathbf{n}}) \rightarrow T \int \frac{d\omega}{2\pi} f(\omega). \quad (4.66)$$

Thus, the information conveyed by observations on a (large) window of time becomes

$$I[y(0 < t < T) \rightarrow x(0 < t < T)] \rightarrow \frac{T}{2} \int \frac{d\omega}{2\pi} \log_2[1 + SNR(\omega)] \text{ bits}. \quad (4.67)$$

We see that the information gained is proportional to the time of our observations, so it makes sense to define an information rate:

$$R_{\text{info}} \equiv \lim_{T \rightarrow \infty} \frac{1}{T} \cdot I[y(0 < t < T) \rightarrow x(0 < t < T)] \quad (4.68)$$

$$= \frac{1}{2} \int \frac{d\omega}{2\pi} \log_2[1 + \text{SNR}(\omega)] \text{ bits/sec.} \quad (4.69)$$

Note that in all these equations, integrals over frequency run over both positive and negative frequencies; if the signals are sampled at points in time spaced by  $\tau_0$  then the maximum (Nyquist) frequency is  $|\omega|_{\text{max}} = \pi/\tau_0$ .

[This requires some more discussion. What we have done is to calculate the rate at which  $y(t)$  provides information about  $x(t)$ , in bits. Notice that information is being transmitted even though there is noise. When we think about bits, we think about answering yes/no questions. Is it possible that this information can be used to (correctly!) answer yes/no questions about  $x(t)$ , even though there is noise, so that relationship between  $y(t)$  and  $x(t)$  has an element of randomness? It turns out that this actually works—it is possible to transmit an average of  $R_{\text{info}}$  yes/no bits of knowledge *without error* as long as these are correctly encoded in  $x(t)$ . This last point, that the bits need to be correctly encoded, is the beginning of a huge field of error-correcting codes. The information rate has a meaning even if we don't talk about these codes, but evidently it has even more meaning if we do. But this is a long detour, and I am not sure how firmly we can connect back to real biological phenomena. So, for now I'll leave it out ... ]

**Problem 44: How long to look?** We know that when we integrate for longer times we can suppress the effects of noise and hence presumably gain more information. Usually we would say that the benefits of integration are cut off by the fact that the signals we are looking at will change. But once we think about information transmission there is another possibility—perhaps we would be better off using the same time to look at something new, rather than getting a more accurate view of something we have already seen. To address this possibility, let's consider the following simple model. We look at one thing for a time  $\tau$ , and then jump to something completely new. Given that we integrate for  $\tau$ , we achieve some signal-to-noise ratio which we'll call  $S(\tau)$ .

(a.) Explain why, in this simple model, the rate at which we gain information is

$$R_{\text{info}}(\tau) = \frac{1}{\tau} \log_2[1 + S(\tau)]. \quad (4.70)$$

How does the assumption that we 'jump to something completely new' enter into the justification of this formula?

(b.) To make progress we need a model for  $S(\tau)$ . Since this is the signal-to-noise ratio let's start with the signal. Suppose that inputs are given by  $x$ , and the output is  $y$ . At  $t = 0$ , the value of  $y$  is set to zero, and after that our sensory responds to its inputs according to a simple differential equation

$$\tau_0 \frac{dy}{dt} = -y + x. \quad (4.71)$$

Show that  $y(\tau) = x[1 - \exp(-\tau/\tau_0)]$ . Now for the noise, suppose that  $\eta_{\text{eff}}(t)$  has a correlation function

$$\langle \eta_{\text{eff}}(t) \eta_{\text{eff}}(t') \rangle = \sigma_0^2 e^{-|t-t'|/\tau_c}. \quad (4.72)$$

Show that if we average the noise over a window of duration  $\tau$ , then the variance

$$\sigma^2(\tau) \equiv \left\langle \left[ \frac{1}{\tau} \int_0^\tau dt \eta_{\text{eff}}(t) \right]^2 \right\rangle \approx \sigma_0^2 \quad (\tau \ll \tau_0) \quad (4.73)$$

$$\approx \frac{2\sigma_0^2 \tau_c}{\tau} \quad (\tau \gg \tau_0). \quad (4.74)$$

Give a more general analytic expression for  $\sigma^2(\tau)$ . Put these factors together to get an expression for  $S(\tau) = y^2(\tau)/\sigma^2(\tau)$ . To keep things simple, you can assume that the time scale which determines the response to inputs is the same as that which determines the correlations in the noise, so that  $\tau_c = \tau_0$ .

(c.) Hopefully you can show from your results in [b] that  $S(\tau \gg \tau_0) \propto \tau$ . This corresponds to our intuition that signal-to-noise ratios grow with averaging time because we beat down the noise, not worrying about the possibility that the signal itself will change. What happens for  $\tau \ll \tau_0$ ?

(d.) Suppose that  $\tau_0$  is very small, so that all “reasonable” values of  $\tau \gg \tau_0$ . Then, from [c],  $S(\tau) = A\tau$ , with  $A$  a constant. With this assumption, plot  $R_{\text{info}}(\tau)$ ; show that with proper choice of units, you don't need to know the value of  $A$ . What value of  $\tau$  maximizes the information rate? Is this consistent with the assumption that  $\tau \gg \tau_0$ ?

(e.) In general, the maximum information is found at the point where  $dR_{\text{info}}/d\tau = 0$ . Show that this condition can be rewritten as a relationship between the signal-to-noise ratio and its logarithmic derivative,  $z = d \ln S(\tau)/d \ln \tau$ . From your previous results, what can you say about the possible values of  $z$  as  $\tau$  is varied? Use this to bound  $S(\tau)$  at the point of maximum  $R_{\text{info}}$ . What does this say about the compromise between looking carefully at one thing and jumping to something new?

(f.) How general can you make the conclusions that you draw in [e]?

The Gaussian channel gives us the opportunity to explore the way in which noise limits information transmission. Imagine that we have measured the spectrum of the effective noise,  $N_{\text{eff}}(\omega)$ . By changing the spectrum of input signals,  $S(\omega)$ , we can change the rate of information transmission. Can we maximize this information rate? Clearly this problem is not well posed without some constraints: if we are allowed just to increase the amplitude of

the signal—multiply the spectrum by a large constant—then we can always increase information transmission. We need to study the optimization of information rate with some fixed ‘dynamic range’ for the signals. A simple example, considered by Shannon at the outset, is to fix the total variance of the signal [Shannon 1949], which is the same as fixing the integral of the spectrum. We can motivate this constraint by noting that if the signal is a voltage and we have to drive this signal through a resistive element, then the variance is proportional to the mean power dissipation. Alternatively, it might be easy to measure the variance of the signals that we are interested in (as for the visual signals in the example below), and then the constraint is empirical.

So the problem we want to solve is maximizing  $R_{\text{info}}$  while holding  $\langle x^2 \rangle$  fixed. As before, we introduce a Lagrange multiplier and maximize a new function

$$\tilde{R} = R_{\text{info}} - \lambda \langle x^2 \rangle \quad (4.75)$$

$$= \frac{1}{2} \int \frac{d\omega}{2\pi} \log_2 \left[ 1 + \frac{S_x(\omega)}{N_{\text{eff}}(\omega)} \right] - \lambda \int \frac{d\omega}{2\pi} S_x(\omega). \quad (4.76)$$

The value of the function  $S_x(\omega)$  at each frequency contributes independently, so it is easy to compute the functional derivatives,

$$\frac{\delta \tilde{R}}{\delta S_x(\omega)} = \frac{1}{2 \ln 2} \cdot \frac{1}{1 + S_x(\omega)/N_{\text{eff}}(\omega)} \cdot \frac{1}{N_{\text{eff}}(\omega)} - \lambda, \quad (4.77)$$

and of course the optimization condition is  $\delta \tilde{R} / \delta S_x(\omega) = 0$ . The result is that

$$S_x(\omega) + N_{\text{eff}}(\omega) = \frac{1}{2\lambda \ln 2}. \quad (4.78)$$

Thus the optimal choice of the signal spectrum is one which makes the sum of signal and (effective) noise equal to white noise! This, like the fact that information is maximized by a Gaussian signal, is telling us that efficient information transmission occurs when the received signals are as random as possible given the constraints. Thus an attempt to look for structure in an optimally encoded signal (say, deep in the brain) will be frustrating.

In general, complete whitening as suggested by Eq. (4.78) can't be achieved at all frequencies, since if the system has finite time resolution (for example) the effective noise grows without bound at high frequencies. Thus the full solution is to have the spectrum determined by Eq. (4.78) everywhere that the spectrum comes out to a positive number, and then to set the spectrum equal to zero outside this range. If we think of the effective

noise spectrum as a landscape with valleys, the condition for optimizing information transmission corresponds to filling the valleys with water; the total volume of water is the variance of the signal.

**Problem 45: Whitening.** Consider a system that responds linearly to a signal  $s(t)$ , with added noise  $\eta(t)$ :

$$x(t) = \int d\tau F(\tau)s(t - \tau) + \eta(t). \quad (4.79)$$

Assume that the noise is Gaussian and white, with power spectrum  $\mathcal{N}_0$ , so that

$$\langle \eta(t)\eta(t') \rangle = \mathcal{N}_0\delta(t - t'). \quad (4.80)$$

For simplicity, assume that the signal  $s(t)$  is Gaussian, with a power spectrum  $S(\omega)$ ,

$$\langle s(t)s(t') \rangle = \int \frac{d\omega}{2\pi} S(\omega) \exp[-i\omega(t - t')]. \quad (4.81)$$

(a.) Write an expression for the rate  $R_{\text{info}}$  at which the observable  $x(t)$  provides information about the signal  $s(t)$ .

(b.) The variance of the variable  $x(t)$  is not well defined. Why? Consider just the component of  $x(t)$  that comes from the signal  $s(t)$ , that is Eq (4.79) but with  $\eta = 0$ . Find an expression for the variance of this “output signal.”

(c.) Consider the problem of maximizing  $R_{\text{info}}$  by adjusting the filter  $F(\tau)$ . Obviously the information transmission is larger if  $F$  is larger, so to make the problem well posed assume that the variance of the output signal (from [b]) is fixed. Show that this variational problem can be solved explicitly for  $|\tilde{F}(\omega)|^2$ , where  $\tilde{F}(\omega)$  is the Fourier transform of the filter  $F(\tau)$ . Can you explain intuitively why only the modulus, and not the phase, of  $\tilde{F}(\omega)$  is relevant here?

(d.) Find the limiting form of the optimal filter as the noise become small. What does this filter do to the input signal? Explain why this makes sense. Saying that “noise is small” is slightly strange, since  $\mathcal{N}_0$  has units. Give a more precise criterion for your small noise limit be valid.

(e.) Consider the case of an input with exponentially decaying correlations, so that

$$S(\omega) = \frac{2\langle s^2 \rangle \tau_c}{1 + (\omega\tau_c)^2}, \quad (4.82)$$

where  $\tau_c$  is the correlation time. Find the optimal filter in this case, and use this to evaluate the maximum value of  $R_{\text{info}}$  as a function of the output signal variance. You should check that your results for  $R_{\text{info}}$ , which should be in bits/s, are independent of the units used for the output variance and the noise power spectrum. Contrast your result with what would happen if  $|\tilde{F}(\omega)|$  were flat as a function of frequency, so that there was no real filtering (just a multiplication so that the output signal variance comes out right). How much can one gain by building the right filter?

These ideas have been used to characterize information transmission across the first synapse in the fly's visual system [de Ruyter van Steveninck and Laughlin 1996]. We have seen these data before, in thinking about how the precision of photon counting changes as the background light intensity increases. Recall that, over a reasonable dynamic range of intensity variations, de Ruyter van Steveninck and Laughlin found that the average voltage response of the photoreceptor cell is related linearly to the intensity or contrast in the movie, and the noise or variability  $\delta V(t)$  is governed by a Gaussian distribution of voltage fluctuations around the average:

$$V(t) = V_{\text{DC}} + \int dt' T(t-t') C(t') + \delta V(t). \quad (4.83)$$

This (happily) is the problem we have just analyzed.

As before, we think of the noise in the response as being equivalent to noise  $\delta C_{\text{eff}}(t)$  that is added to the movie itself,

$$V(t) = V_{\text{DC}} + \int dt' T(t-t') [C(t') + \delta C_{\text{eff}}(t')]. \quad (4.84)$$

Since the fluctuations have a Gaussian distribution, they can be characterized completely by their power spectrum  $N_C^{\text{eff}}(\omega)$ , which measures the variance of the fluctuations that occur at different frequencies,

$$\langle \delta C_{\text{eff}}(t) \delta C_{\text{eff}}(t') \rangle = \int \frac{d\omega}{2\pi} N_C^{\text{eff}}(\omega) \exp[-i\omega(t-t')]. \quad (4.85)$$

There is a minimum level of this effective noise set by the random arrival of photons (shot noise). The photon noise is white if expressed as  $N_C^{\text{eff}}(\omega)$ , although it makes a nonwhite contribution to the voltage noise. As we have discussed, over a wide range of background light intensities and frequencies, the fly photoreceptors have effective noise levels that reach the limit set by photon statistics. At high frequencies there is excess noise beyond the physical limit, and this excess noise sets the time resolution of the system.

The power spectrum of the effective noise tells us, ultimately, what signals the photoreceptor can and cannot transmit. How do we turn these measurements into bits? One approach is to assume that the fly lives in some particular environment, and then calculate how much information the receptor cell can provide about this particular environment. But to characterize the cell itself, we might ask a different question: in principle how much information can the cell transmit? To answer this question we are allowed to shape the statistical structure of the environment so as to make the best use of the receptor (the opposite, presumably, of what happens in

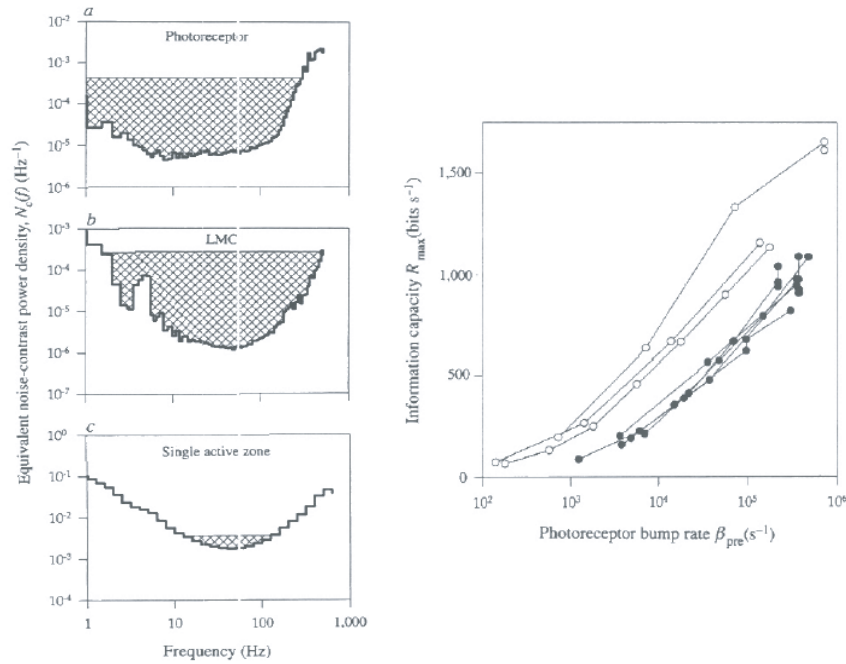


Figure 4.1: At left, the effective contrast noise levels in a single photoreceptor cell, a single LMC (the second order cell) and the inferred noise level for a single active zone of the synapse from photoreceptor to LMC. The hatching shows the signal spectra required to whiten the total output over the largest possible range while maintaining the input contrast variance  $\langle C^2 \rangle = 0.1$ , as discussed in the text. At right, the resulting information capacities as a function of the photon counting rates in the photoreceptors. From [de Ruyter van Steveninck & Laughlin 1996a].

evolution!). This is just the optimization discussed above, so it is possible to turn the measurements on signals and noise into estimates of the information capacity of these cells. This was done both for the photoreceptor cells and for the large monopolar cells that receive direct synaptic input from a group of six receptors. From measurements on natural scenes the mean square contrast signal was fixed at  $\langle C^2 \rangle = 0.1$ . Results are shown in Fig 4.1.

The first interesting feature of the results is the scale: individual neurons are capable of transmitting well above 1000 bits per second. This does not mean that this capacity is used under natural conditions, but rather speaks to the precision of the mechanisms underlying the detection and transmission of signals in this system. Second, information capacity continues to increase

as the level of background light increases: noise due to photon statistics is less important in brighter lights, and this reduction of the physical limit actually improves the performance of the system even up to very high photon counting rates, indicating once more that the physical limit is relevant to the real performance. Third, we see that the information capacity as a function of photon counting rate is shifted along the counting rate axis as we go from photoreceptors to LMCs, and this corresponds (quite accurately!) to the fact that LMCs integrate signals from six photoreceptors and thus act as if they captured photons at six times higher rate. Finally, in the large monopolar cells information has been transmitted across a synapse, and in the process is converted from a continuous voltage signal into discrete events corresponding to the release of neurotransmitter vesicles at the synapse. As a result, there is a new limit to information transmission that comes from viewing the large monopolar cell as a “vesicle counter.”

If every vesicle makes a measurable, deterministic contribution to the cell’s response (a generous assumption), then the large monopolar cell’s response is equivalent to reporting how many vesicles are counted in a small window of time corresponding to the photoreceptor time resolution. We don’t know the distribution of these counts, but we can estimate (from other experiments, with uncertainty) the mean count, and we know that there is a maximum entropy for any count distribution once we fix the mean [need to point to results in another section of the course]. No mechanism at the synapse can transmit more information than this limit. Remarkably, the fly operates within a factor of two of this limit, and the agreement might be even better but for uncertainties in the vesicle counting rate [Rieke et al 1997].

This example, from the first synapse in fly vision, provides evidence that the visual system really has constructed a transformation of light intensity into transmembrane voltage that is efficient in the sense defined by information theory. In fact there is more to the analysis of even this one synapse, as we shall see. But, armed with some suggestive results, let’s go on ... .

