Some background on information theory

These notes are based on a section of a full biophysics course which I teach at Princeton, aimed at PhD students in Physics. In the current organization, this section includes an account of Laughlin's ideas of optimization in the fly retina, as well as some pointers to topics in the next lecture. Gradually the whole course is being turned into a book, to be published by Princeton University Press. I hope this extract is useful in the context of our short course, and of course I would be grateful for any feedback.

The generation of physicists who turned to biological phenomena in the wake of quantum mechanics noted that to understand life one has to understand not just the flow of energy (as in inanimate systems) but also the flow of information. There is, of course, some difficulty in translating the colloquial notion of information into something mathematically precise. Indeed, almost all statistical mechanics textbooks note that the entropy of a gas measures our lack of information about the microscopic state of the molecules, but often this connection is left a bit vague or qualitative. Shannon proved a theorem that makes the connection precise [Shannon 1948]: entropy is the unique measure of available information consistent with certain simple and plausible requirements. Further, entropy also answers the practical question of how much space we need to use in writing down a description of the signals or states that we observe. This leads to a notion of *efficient representation*, and in this section of the course we'll explore the possibility that biological systems in fact form efficient representations of relevant information.

I. ENTROPY AND INFORMATION

Two friends, Max and Allan, are having a conversation. In the course of the conversation, Max asks Allan what he thinks of the headline story in this morning's newspaper. We have the clear intuitive notion that Max will 'gain information' by hearing the answer to his question, and we would like to quantify this intuition. Following Shannon's reasoning, we begin by assuming that Max knows Allan very well. Allan speaks very proper English, being careful to follow the grammatical rules even in casual conversation. Since they have had many political discussions Max has a rather good idea about how Allan will react to the latest news. Thus Max can make a list of Allan's possible responses to his question, and he can assign probabilities to each of the answers. From this list of possibilities and probabilities we can compute an entropy, and this is done in exactly the same way as we compute the entropy of a gas in statistical mechanics or thermodynamics: If the probability of the n^{th} possible response is p_{n}, then the entropy is

$$S = -\sum_{\mathrm{n}} p_{\mathrm{n}} \log_2 p_{\mathrm{n}} \text{ bits}. \tag{1}$$

The entropy S measures Max's uncertainty about what Allan will say in response to his question. Once Allan gives his answer, all this uncertainty is removed—one of the responses occurred, corresponding to $p = 1$, and all the others did not, corresponding to $p = 0$—so the entropy is reduced to zero. It is appealing to equate this reduction in our uncertainty with the information we gain by hearing Allan's answer. Shannon proved that this is not just an interesting analogy; it is the *only* definition of information that conforms to some simple constraints.

A. Shannon's uniqueness theorem

To start, Shannon assumes that the information gained on hearing the answer can be written as a function of the probabilities p_{n}.[1] Then if all N possible answers are equally likely the information gained should be a monotonically increasing function of N. The next constraint is that if our question consists of two parts, and if these two parts are entirely independent of one another, then we should be able to write the total information gained as the sum of the information gained in response to each of the two subquestions. Finally, more general multipart questions can be thought of as branching trees, where the answer to each successive part of the question provides some further refinement of the probabilities; in this case we should be able to write the total information gained as the weighted sum of the information gained at each branch point. Shannon proved that the only function of the $\{p_{\mathrm{n}}\}$ consistent with these three postulates—monotonicity, independence, and branching—is the entropy S, up to a multiplicative constant.

To prove Shannon's theorem we start with the case where all N possible answers are equally likely. Then the information must be a function of N, and let this function be $f(N)$. Consider the special case $N = k^m$. Then we can think of our answer—one out of N possibilities—as being given in m independent parts, and in each part we must be told one of k equally likely possibilities.[2]

---

[1] In particular, this 'zeroth' assumption means that we must take seriously the notion of enumerating the possible answers. In this framework we cannot quantify the information that would be gained upon hearing a previously unimaginable answer to our question.

[2] If $N = k^m$, then we can specify each possible answer by an m–

But we have assumed that information from independent questions and answers must add, so the function $f(N)$ must obey the condition

$$f(k^m) = mf(k). \tag{2}$$

It is easy to see that $f(N) \propto \log N$ satisfies this equation. To show that this is the unique solution, Shannon considers another pair of integers ℓ and n such that

$$k^m \leq \ell^n \leq k^{m+1}, \tag{3}$$

or, taking logarithms,

$$\frac{m}{n} \leq \frac{\log \ell}{\log k} \leq \frac{m}{n} + \frac{1}{n}. \tag{4}$$

Now because the information measure $f(N)$ is monotonically increasing with N, the ordering in Eq. (3) means that

$$f(k^m) \leq f(\ell^n) \leq f(k^{m+1}), \tag{5}$$

and hence from Eq. (2) we obtain

$$mf(k) \leq nf(\ell) \leq (m+1)f(k). \tag{6}$$

Dividing through by $nf(k)$ we have

$$\frac{m}{n} \leq \frac{f(\ell)}{f(k)} \leq \frac{m}{n} + \frac{1}{n}, \tag{7}$$

which is very similar to Eq. (4). The trick is now that with k and ℓ fixed, we can choose an arbitrarily large value for n, so that $1/n = \epsilon$ is as small as we like. Then Eq. (4) is telling us that

$$\left| \frac{m}{n} - \frac{\log \ell}{\log k} \right| < \epsilon, \tag{8}$$

and hence Eq. (7) for the function $f(N)$ can similarly be rewritten as

$$\left| \frac{m}{n} - \frac{f(\ell)}{f(k)} \right| < \epsilon, \text{ or} \tag{9}$$

$$\left| \frac{f(\ell)}{f(k)} - \frac{\log \ell}{\log k} \right| \leq 2\epsilon, \tag{10}$$

so that $f(N) \propto \log N$ as promised.[3]

We are not quite finished, even with the simple case of N equally likely alternatives, because we still have an arbitrary constant of proportionality. We recall that the same issue arises is statistical mechanics: what are the units of entropy? In a physical chemistry course you might learn that entropy is measured in "entropy units," with the property that if you multiply by the absolute temperature (in Kelvin) you obtain an energy in units of calories per mole; this happens because the constant of proportionality is chosen to be the gas constant R, which refers to Avogadro's number of molecules.[4] In physics courses entropy is often defined with a factor of Boltzmann's constant k_B, so that if we multiply by the absolute temperature we again obtain an energy (in Joules) but now per molecule (or per degree of freedom), not per mole. In fact many statistical mechanics texts take the sensible view that temperature itself should be measured in energy units—that is, we should always talk about the quantity k_BT, not T alone—so that the entropy, which after all measures the number of possible states of the system, is dimensionless. Any dimensionless proportionality constant can be absorbed by choosing the base that we use for taking logarithms, and in information theory it is conventional to choose base two. Finally, then, we have $f(N) = \log_2 N$. The units of this measure are called *bits,* and one bit is the information contained in the choice between two equally likely alternatives.

Ultimately we need to know the information conveyed in the general case where our N possible answers all have unequal probabilities. To make a bridge to this general problem from the simple case of equal probabilities, consider the situation where all the probabilities are rational, that is

$$p_{\rm n} = \frac{k_{\rm n}}{\sum_{\rm m} k_{\rm m}}, \tag{11}$$

where all the $k_{\rm n}$ are integers. It should be clear that if we can find the correct information measure for rational $\{p_{\rm n}\}$ then by continuity we can extrapolate to the general case; the trick is that we can reduce the case of rational probabilities to the case of equal probabilities. To do this, imagine that we have a total of $N_{\rm total} = \sum_{\rm m} k_{\rm m}$ possible answers, but that we have organized these into N groups, each of which contains $k_{\rm n}$ possibilities. If we specified the full answer, we would first tell which group it was in, then tell which of the $k_{\rm n}$ possibilities was realized. In this two step process, at the first step we get the information we are really looking for—which of the N groups are we in—and so the information in

place ('digit' is the wrong word for $k \neq 10$) number in base k. The m independent parts of the answer correspond to specifying the number from 0 to $k-1$ that goes in each of the m places.

[3] If we were allowed to consider $f(N)$ as a continuous function, then we could make a much simpler argument. But, strictly speaking, $f(N)$ is defined only at integer arguments.

[4] Whenever I read about entropy units (or calories, for that matter) I imagine that there was some great congress on units at which all such things were supposed to be standardized. Of course every group has its own favorite nonstandard units. Perhaps at the end of some long negotiations the chemists were allowed to keep entropy units in exchange for physicists continuing to use electron Volts.

the first step is our unknown function,

$$I_1 = I(\{p_n\}). \tag{12}$$

At the second step, if we are in group n then we will gain $\log_2 k_n$ bits, because this is just the problem of choosing from k_n equally likely possibilities, and since group n occurs with probability p_n the *average* information we gain in the second step is

$$I_2 = \sum_n p_n \log_2 k_n. \tag{13}$$

But this two step process is not the only way to compute the information in the enlarged problem, because, by construction, the enlarged problem is just the problem of choosing from $N_{\rm total}$ equally likely possibilities. The two calculations have to give the same answer, so that

$$I_1 + I_2 = \log_2(N_{\rm total}), \tag{14}$$

$$I(\{p_n\}) + \sum_n p_n \log_2 k_n = \log_2(\sum_m k_m). \tag{15}$$

Rearranging the terms, we find

$$I(\{p_n\}) = -\sum_n p_n \log_2 \left(\frac{k_n}{\sum_m k_m} \right) \tag{16}$$

$$= -\sum_n p_n \log_2 p_n. \tag{17}$$

Again, although this is worked out explicitly for the case where the p_n are rational, it must be the general answer if the information measure is continuous. So we are done: the information obtained on hearing the answer to a question is measured uniquely by the entropy of the distribution of possible answers.

If we phrase the problem of gaining information from hearing the answer to a question, then it is natural to think about a discrete set of possible answers. On the other hand, if we think about gaining information from the acoustic waveform that reaches our ears, then there is a continuum of possibilities. Naively, we are tempted to write

$$S_{\rm continuum} = -\int dx P(x) \log_2 P(x), \tag{18}$$

or some multidimensional generalization. The difficulty, of course, is that probability distributions for continuous variables [like $P(x)$ in this equation] have units—the distribution of x has units inverse to the units of x—and we should be worried about taking logs of objects that have dimensions. Notice that if we wanted to compute a difference in entropy between two distributions, this problem would go away. This is a hint that only entropy differences are going to be important.[5]

[5] The problem of defining the entropy for continuous variables is familiar in statistical mechanics. In the simple example of an ideal gas in a finite box, we know that the quantum version of the problem has a discrete set of states, so that we can compute the entropy of the gas as a sum over these states. In the limit that the box is large, sums can be approximated as integrals, and if the temperature is high we expect that quantum effects are negligible and one might naively suppose that Planck's constant should disappear from the results; we recall that this is not quite the case. Planck's constant has units of momentum times position, and so is an elementary area for each pair of conjugate position and momentum variables in the classical phase space; in the classical limit the entropy becomes (roughly) the logarithm of the occupied volume in phase space, but this volume is measured in units of Planck's constant. If we had tried to start with a classical formulation (as did Boltzmann and Gibbs, of course) then we would find ourselves with the problems of Eq. (18), namely that we are trying to take the logarithm of a quantity with dimensions. If we measure phase space volumes in units of Planck's constant, then all is well. The important point is that the problems with defining a purely classical entropy do *not* stop us from calculating entropy differences, which are observable directly as heat flows, and we shall find a similar situation for the information content of continuous ("classical") variables.

B. Writing down the answers

In the simple case where we ask a question and there are exactly $N = 2^m$ possible answers, all with equal probability, the entropy is just m bits. But if we make a list of all the possible answers we can label each of them with a distinct m–bit binary number: to specify the answer all I need to do is write down this number. Note that the answers themselves can be very complex: different possible answers could correspond to lengthy essays, but the number of pages required to write these essays is irrelevant. If we agree in advance on the set of possible answers, all I have to do in answering the question is to provide a unique label. If we think of the label as a 'codeword' for the answer, then in this simple case the length of the codeword that represents the $n^{\rm th}$ possible answer is given by $\ell_n = -\log_2 p_n$, and the average length of a codeword is given by the entropy.

It will turn out that the equality of the entropy and the average length of codewords is much more general than our simple example. Before proceding, however, it is important to realize that the entropy is emerging as the answer to two very different questions. In the first case we wanted to quantify our intuitive notion of gaining information by hearing the answer to a question. In the second case, we are interested in the problem of *representing* this answer in the smallest possible space. It is quite remarkable that the only way of quantifying how much we learn by hearing the answer to a question is to measure how much space is required to write down the answer.

Clearly these remarks are interesting only if we can treat more general cases. Let us recall that in statistical mechanics we have the choice of working with a microcanonical ensemble, in which an ensemble of systems is

distributed uniformly over states of fixed energy, or with a canonical ensemble, in which an ensemble of systems is distributed across states of different energies according to the Boltzmann distribution. The microcanonical ensemble is like our simple example with all answers having equal probability: entropy really is just the log of the number of possible states. On the other hand, we know that in the thermodynamic limit there is not much difference between the two ensembles. This suggests that we can recover a simple notion of representing answers with codewords of length $\ell_{\rm n} = -\log_2 p_{\rm n}$ provided that we can find a suitbale analog of the thermodynamic limit.

Imagine that instead of asking a question once, we ask it many times. As an example, every day we can ask the weatherman for an estimate of the temperature at noon the next day. Now instead of trying to represent the answer to one question we can try to represent the whole stream of answers collected over a long period of time. Thus instead of a possible answer being labelled n, possible sequences of answers are labelled by ${\rm n}_1{\rm n}_2\cdots{\rm n}_N$. Of course these sequences have probabilites $P({\rm n}_1{\rm n}_2\cdots{\rm n}_N)$, and from these probabilities we can compute an entropy that must depend on the length of the sequence,

$$S(N) = -\sum_{{\rm n}_1}\sum_{{\rm n}_2}\cdots\sum_{{\rm n}_N} P({\rm n}_1{\rm n}_2\cdots{\rm n}_N)\log_2 P({\rm n}_1{\rm n}_2\cdots{\rm n}_N). \tag{19}$$

Notice we are *not* assuming that successive questions have independent answers, which would correspond to $P({\rm n}_1{\rm n}_2\cdots{\rm n}_N) = \prod_{i=1}^N p_{{\rm n}_i}$.

Now we can draw on our intuition from statistical mechanics. The entropy is an extensive quantity, which means that as N becomes large the entropy should be proportional to N; more precisely we should have

$$\lim_{N\to\infty}\frac{S(N)}{N} = \mathcal{S}, \tag{20}$$

where $\mathcal{S}$ is the entropy density for our sequence in the same way that a large volume of material has a well defined entropy per unit volume.

The equivalence of ensembles in the thermodynamic limit means that having unequal probabilities in the Boltzmann distribution has almost no effect on anything we want to calculate. In particular, for the Boltzmann distribution we know that, state by state, the log of the probability is the energy and that this energy is itself an extensive quantity. Further we know that (relative) fluctuations in energy are small. But if energy is log probability, and relative fluctuations in energy are small, this must mean that almost all the states we actually observe have log probabilities which are the same. By analogy, all the long sequences of answers must fall into two groups: those with $-\log_2 P \approx N\mathcal{S}$, and those with $P \approx 0$. Now this is all a bit sloppy, but it is the right idea: if we are willing to think about long sequences or streams of data, then the equivalence of ensembles tells us that 'typical' sequences are uniformly distributed over $\mathcal{N} \approx 2^{N\mathcal{S}}$ possibilities, and that this appproximation becomes more and more accurate as the length N of the sequences becomes large.

The idea of typical sequences, which is the information theoretic version of a thermodynamic limit, is enough to tell us that our simple arguments about representing answers by binary numbers ought to work on average for long sequences of answers. We will have to work significantly harder to show that this is really the smallest possible representation. An important if obvious consequence is that if we have many rather unlikekly answers (rather than fewer more likely answers) then we need more space to write the answers down. More profoundly, this turns out to be true answer by answer: to be sure that long sequences of answers take up as little space as possible, we need to use an average of $\ell_{\rm n} = -\log_2 p_{\rm n}$ bits to represent each individual answer n. Thus answers which are more surprising require more space to write down.

C. Entropy lost and information gained

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 $P(D)$. 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:

$$
\begin{aligned}
I(D \to W) &= S(W) - \sum_D P(D) S(W|D) && (21)\\
&= -\sum_W P(W) \log_2 P(W) - \sum_D P(D) \left[-\sum_W P(W|D) \log_2 P(W|D) \right] && (22)\\
&= -\sum_W \sum_D P(W,D) \log_2 P(W) + \sum_W \sum_D P(W|D) P(D) \log_2 P(W|D) && (23)\\
&= -\sum_W \sum_D P(W,D) \log_2 P(W) + \sum_W \sum_D P(W,D) \log_2 P(W|D) && (24)\\
&= \sum_W \sum_D P(W,D) \log_2 \left[\frac{P(W|D)}{P(W)} \right] && (25)\\
&= \sum_W \sum_D P(W,D) \log_2 \left[\frac{P(W,D)}{P(W)P(D)} \right]. && (26)
\end{aligned}
$$

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 $I(D \to W)$.

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

$$
\begin{aligned}
I(D;W) &= S(W) - \sum_D P(D) S(W|D) && (27)\\
&= S(D) - \sum_W P(W) S(D|W). && (28)
\end{aligned}
$$

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

$$
\begin{aligned}
I(D;W) &\leq S(W) && (29)\\
I(D;W) &\leq S(D). && (30)
\end{aligned}
$$

The first equation tells us that by observing D we cannot learn more about the world then 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.[6] The second equation, 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[7] thus limits how much

[6] This is not quite true. 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.

[7] In the same way that we speak about the entropy of a gas I will often speak about the entropy of a variable or the entropy of a

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).

D. Optimizing input/output relations

These ideas are enough to get started on "designing" some simple neural processes [Laughlin 1981]. Imagine that a neuron is responsible for representing a single number such as the light intensity $\mathcal{I}$ averaged over a small patch of the retina (don't worry about time dependence). Assume that this signal will be represented by a continuous voltage V, which is true for the first stages of processing in vision, as we have seen in the first section of the course. The information that the voltage provides about the intensity is

$$I(V \to \mathcal{I}) = \int d\mathcal{I} \int dV\, P(V,\mathcal{I}) \log_2 \left[\frac{P(V,\mathcal{I})}{P(V)P(\mathcal{I})} \right] \quad (31)$$

$$= \int d\mathcal{I} \int dV\, P(V,\mathcal{I}) \log_2 \left[\frac{P(V|\mathcal{I})}{P(V)} \right]. \quad (32)$$

The conditional distribution $P(V|\mathcal{I})$ describes the process by which the neuron responds to its input, and so this is what we should try to "design."

Let us suppose that the voltage is on average a nonlinear function of the intensity, and that the dominant source of noise is additive (to the voltage), independent of light intensity, and small compared with the overall dynamic range of the cell:

$$V = g(\mathcal{I}) + \xi, \quad (33)$$

with some distribution $P_{\rm noise}(\xi)$ for the noise. Then the conditional distribution

$$P(V|\mathcal{I}) = P_{\rm noise}(V - g(\mathcal{I})), \quad (34)$$

and the entropy of this conditional distribution can be written as

$$S_{\rm cond} = -\int dV\, P(V|\mathcal{I}) \log_2 P(V|\mathcal{I}) \quad (35)$$

$$= -\int d\xi\, P_{\rm noise}(\xi) \log_2 P_{\rm noise}(\xi). \quad (36)$$

Note that this is a constant, independent both of the light intensity and of the nonlinear input/output relation $g(\mathcal{I})$. This is useful because we can write the information as a difference between the total entropy of the output variable V and this conditional or noise entropy, as in Eq. (28):

$$I(V \to \mathcal{I}) = -\int dV P(V) \log_2 P(V) - S_{\rm cond}. \quad (37)$$

With $S_{\rm cond}$ constant independent of our 'design,' maximizing information is the same as maximizing the entropy of the distribution of output voltages. Assuming that there are maximum and minimum values for this voltage, but no other constraints, then the maximum entropy distribution is just the uniform distribution within the allowed dynamic range. But if the noise is small it doesn't contribute much to broadening $P(V)$ and we calculate this distribution as if there were no noise, so that

$$P(V)dV = P(\mathcal{I})d\mathcal{I}, \quad (38)$$

$$\frac{dV}{d\mathcal{I}} = \frac{1}{P(V)} \cdot P(\mathcal{I}). \quad (39)$$

Since we want to have $V = g(\mathcal{I})$ and $P(V) = 1/(V_{\rm max} - V_{\rm min})$, we find

$$\frac{dg(\mathcal{I})}{d\mathcal{I}} = (V_{\rm max} - V_{\rm min}) P(\mathcal{I}), \quad (40)$$

$$g(\mathcal{I}) = (V_{\rm max} - V_{\rm min}) \int_{\mathcal{I}_{\rm min}}^{\mathcal{I}} d\mathcal{I}' P(\mathcal{I}'). \quad (41)$$

Thus, the optimal input/output relation is proportional to the cumulative probability distribution of the input signals.

The predictions of Eq. (41) are quite interesting. First of all it makes clear that any theory of the nervous system which involves optimizing information transmission or efficiency of representation inevitably predicts that the computations done by the nervous system must be matched to the statistics of sensory inputs (and, presumably, to the statistics of motor outputs as well). Here the matching is simple: in the right units we could just read off the distribution of inputs by looking at the (differentiated) input/output relation of the neuron. Second, this simple model automatically carries some predictions about adaptation to overall light levels. If we live in a world with diffuse light sources that are not directly visible, then the intensity which reaches us at a point is the product of the effective brightness of the source and some local reflectances. As is it gets dark outside the reflectances don't change—these are material properties—and so we expect that the distribution $P(\mathcal{I})$ will look the same except for scaling. Equivalently, if we view the input as the log of the intensity, then to a good approximation $P(\log \mathcal{I})$ just shifts linearly along the $\log \mathcal{I}$ axis

response. In the gas, we understand from statistical mechanics that the entropy is defined not as a property of the gas but as a property of the distribution or ensemble from which the microscopic states of the gas are chosen; similarly we should really speak here about "the entropy of the distribution of observations," but this is a bit cumbersome. I hope that the slightly sloppy but more compact phrasing does not cause confusion.

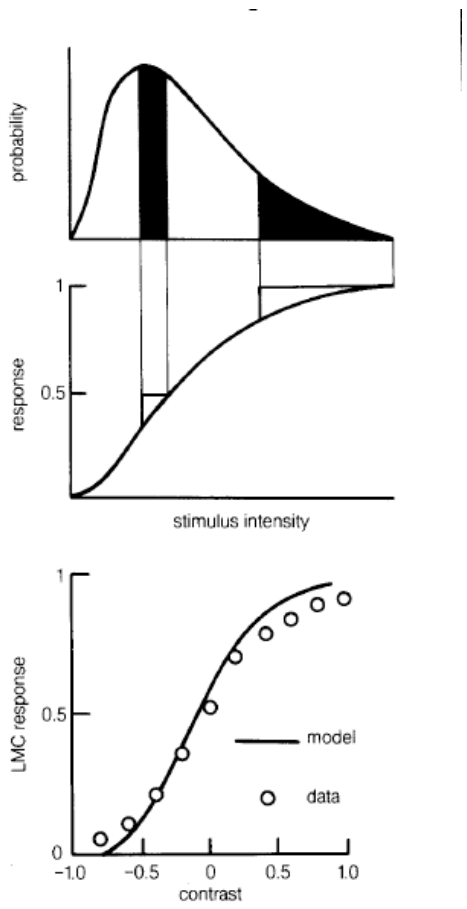

FIG. 1 Matching of input/output relation to the distribution of inputs in the fly large monopolar cells (LMCs). Top, a schematic probability distribution for light intensity. Middle, the optimal input/output relation according to Eq (41), highlighting the fact that equal ranges of output correspond to equal mass under the input distribution. Bottom, measurements of the input/output relation in LMCs compared with these predictions From [Laughlin 1981 & 1987].

as mean light intensity goes up and down. But then the optimal input/output relation $g(\mathcal{I})$ would exhibit a similar invariant shape with shifts along the input axis when expressed as a function of $\log \mathcal{I}$, and this is in rough agreement with experiments on light/dark adaptation in a wide variety of visual neurons. Finally, although obviously a simplified version of the real problem facing even the first stages of visual processing, this calculation does make a quantitative prediction that would be tested if we measure both the input/output relations of early visual neurons and the distribution of light intensities that the animal encounters in nature.

Laughlin made this comparison (25 years ago!) for the fly visual system [Laughlin 1981]. He built an electronic photodetector with aperture and spectral sensitivity matched to those of the fly retina and used his photodetector to scan natural scenes, measuring $P(\mathcal{I})$ as it would appear at the input to these neurons. In parallel he characterized the second order neurons of the fly visual system—the large monopolar cells which receive direct synaptic input from the photoreceptors—by measuring the peak voltage response to flashes of light. The agreement with Eq. (41) was remarkable (Fig 1), especially when we remember that there are no free parameters. While there are obvious open questions (what happened to time dependence?), this is a really beautiful result that inspires us to take these ideas more seriously.

E. Maximum entropy distributions

Since the information we can gain is limited by the entropy, it is natural to ask if we can put limits on the entropy using some low order statistical properties of the data: the mean, the variance, perhaps higher moments or correlation functions, ... In particular, if we can say that the entropy has a maximum value consistent with the observed statistics, then we have placed a firm upper bound on the information that these data can convey.

The problem of finding the maximum entropy given some constraint again is familiar from statistical mechanics: the Boltzmann distribution is the distribution that has the largest possible entropy given the mean energy. More generally, imagine that we have knowledge not of the whole probability distribution $P(D)$ but only of some expectation values,

$$\langle f_{\rm i}\rangle = \sum_D P(D) f_{\rm i}(D), \tag{42}$$

where we allow that there may be several expectation values known (i $= 1, 2, ..., K$). Actually there is one more expectation value that we always know, and this is that the average value of one is one; the distribution is normalized:

$$\langle f_0\rangle = \sum_D P(D) = 1. \tag{43}$$

Given the set of numbers $\{\langle f_0\rangle, \langle f_1\rangle, \cdots, \langle f_K\rangle\}$ as constraints on the probability distribution $P(D)$, we would like to know the largest possible value for the entropy, and we would like to find explicitly the distribution that provides this maximum.

The problem of maximizing a quantity subject to constraints is formulated using Lagrange multipliers. In this case, we want to maximize $S = -\sum P(D)\log_2 P(D)$, so we introduce a function $\tilde{S}$, with one Lagrange multiplier $\lambda_{\rm i}$ for each constraint:

$$\tilde{S}[P(D)] = -\sum_D P(D)\log_2 P(D) - \sum_{\mathrm{i}=0}^{K}\lambda_\mathrm{i}\langle f_\mathrm{i}\rangle \quad (44)$$

$$= -\frac{1}{\ln 2}\sum_D P(D)\ln P(D) - \sum_{\mathrm{i}=0}^{K}\lambda_\mathrm{i}\sum_D P(D) f_\mathrm{i}(D). \quad (45)$$

Our problem is then to find the maximum of the function $\tilde{S}$, but this is easy because the probability for each value of D appears independently. The result is that

$$P(D) = \frac{1}{Z}\exp\left[-\sum_{\mathrm{i}=1}^{K}\lambda_\mathrm{i} f_\mathrm{i}(D)\right], \quad (46)$$

where $Z = \exp(1+\lambda_0)$ is a normalization constant.

There are few more things worth saying about maximum entropy distributions. First, we recall that if the value of D indexes the states n of a physical system, and we know only the expectation value of the energy,

$$\langle E\rangle = \sum_\mathrm{n} P_\mathrm{n} E_\mathrm{n}, \quad (47)$$

then the maximum entropy distribution is

$$P_\mathrm{n} = \frac{1}{Z}\exp(-\lambda E_\mathrm{n}), \quad (48)$$

which is the Boltzmann distribution (as promised). In this case the Lagrange multiplier λ has physical meaning—it is the inverse temperature. Further, the function $\tilde{S}$ that we introduced for convenience is the difference between the entropy and λ times the energy; if we divide through by λ and flip the sign, then we have the energy minus the temperature times the entropy, or the free energy. Thus the distribution which maximizes entropy at fixed average energy is also the distribution which minimizes the free energy.

If we are looking at a magnetic system, for example, and we know not just the average energy but also the average magnetization, then a new term appears in the exponential of the probability distribution, and we can interpret this term as the magnetic field multiplied by the magnetization. More generally, for every order parameter which we assume is known, the probability distribution acquires a term that adds to the energy and can be thought of as a product of the order parameter with its conjugate force. Again, all these remarks should be familiar from a statistical mechanics course.

Probability distributions that have the maximum entropy form of Eq. (46) are special not only because of their connection to statistical mechanics, but because they form what the statisticians call an 'exponential family,' which seems like an obvious name. The important point is that exponential families of distributions are (almost) unique in having sufficient statistics. To understand what this means, consider the following problem: We observe a set of samples $D_1, D_2, \cdots, D_N$, each of which is drawn independently and at random from a distribution $P(D|\{\lambda_\mathrm{i}\})$. Assume that we know the form of this distribution but not the values of the parameters $\{\lambda_\mathrm{i}\}$. How can we estimate these parameters from the set of observations $\{D_\mathrm{n}\}$? Notice that our data set $\{D_\mathrm{n}\}$ consists of N numbers, and N can be very large; on the other hand there typically are a small number $K \ll N$ of parameters λ_i that we want to estimate. Even in this limit, no finite amount of data will tell us the exact values of the parameters, and so we need a probabilistic formulation: we want to compute the distribution of parameters given the data, $P(\{\lambda_\mathrm{i}\}|\{D_\mathrm{n}\})$. We do this using Bayes' rule,

$$P(\{\lambda_\mathrm{i}\}|\{D_\mathrm{n}\}) = \frac{1}{P(\{D_\mathrm{n}\})}\cdot P(\{D_\mathrm{n}\}|\{\lambda_\mathrm{i}\})P(\{\lambda_\mathrm{i}\}), \quad (49)$$

where $P(\{\lambda_\mathrm{i}\})$ is the distribution from which the parameter values themselves are drawn. Then since each datum D_n is drawn independently, we have

$$P(\{D_\mathrm{n}\}|\{\lambda_\mathrm{i}\}) = \prod_{\mathrm{n}=1}^{N} P(D_\mathrm{n}|\{\lambda_\mathrm{i}\}). \quad (50)$$

For probability distributions of the maximum entropy form we can proceed further, using Eq. (46):

$$P(\{\lambda_\mathrm{i}\}|\{D_\mathrm{n}\}) = \frac{1}{P(\{D_\mathrm{n}\})}\cdot P(\{D_\mathrm{n}\}|\{\lambda_\mathrm{i}\})P(\{\lambda_\mathrm{i}\})$$

$$= \frac{P(\{\lambda_\mathrm{i}\})}{P(\{D_\mathrm{n}\})}\prod_{\mathrm{n}=1}^{N} P(D_\mathrm{n}|\{\lambda_\mathrm{i}\}) \quad (51)$$

$$= \frac{P(\{\lambda_i\})}{Z^N P(\{D_n\})} \prod_{n=1}^{N} \exp\left[-\sum_{i=1}^{K} \lambda_i f_i(D_n)\right] \quad (52)$$

$$= \frac{P(\{\lambda_i\})}{Z^N P(\{D_n\})} \exp\left[-N \sum_{i=1}^{K} \lambda_i \frac{1}{N} \sum_{n=1}^{N} f_i(D_n)\right]. \quad (53)$$

We see that *all* of the information that the data points $\{D_n\}$ can give about the parameters λ_i is contained in the average values of the functions f_i over the data set, or the 'empirical means' $\bar{f}_i$,

$$\bar{f}_i = \frac{1}{N} \sum_{n=1}^{N} f_i(D_n). \quad (54)$$

More precisely, the distribution of possible parameter values consistent with the data depends not on all details of the data, but rather only on the empirical means $\{\bar{f}_i\}$,

$$P(\{\lambda_i\}|D_1, D_2, \cdots, D_N) = P(\{\lambda_i\}|\{\bar{f}_j\}), \quad (55)$$

and a consequence of this is the information theoretic statement

$$I(D_1, D_2, \cdots, D_N \to \{\lambda_i\}) = I(\{\bar{f}_j\} \to \{\lambda_i\}). \quad (56)$$

This situation is described by saying that the reduced set of variables $\{\bar{f}_j\}$ constitute *sufficient statistics* for learning the distribution. Thus, for distributions of this form, the problem of compressing N data points into $K << N$ variables that are relevant for parameter estimation can be solved explicitly: if we keep track of the running averages $\bar{f}_i$ we can compress our data as we go along, and we are guaranteed that we will never need to go back and examine the data in more detail. A clear example is that if we know data are drawn from a Gaussian distribution, running estimates of the mean and variance contain all the information available about the underlying parameter values.

The Gaussian example makes it seem that the concept of sufficient statistics is trivial: of course if we know that data are chosen from a Gaussian distribution, then to identify the distribution all we need to do is to keep track of two moments. Far from trivial, this situation is quite unusual. Most of the distributions that we might write down do not have this property—even if they are described by a finite number of parameters, we cannot guarantee that a comparably small set of empirical expectation values captures all the information about the parameter values. If we insist further that the sufficient statistics be additive and permutation symmetric,[8] then it is a theorem that *only* exponential families have sufficient statistics.

The generic problem of information processing, by the brain or by a machine, is that we are faced with a huge quantity of data and must extract those pieces that are of interest to us. The idea of sufficient statistics is intriguing in part because it provides an example where this problem of 'extracting interesting information' can be solved completely: if the points $D_1, D_2, \cdots, D_N$ are chosen independently and at random from some distribution, the only thing which could possibly be 'interesting' is the structure of the distribution itself (everything else is random, by construction), this structure is described by a finite number of parameters, and there is an explicit algorithm for compressing the N data points $\{D_n\}$ into K numbers that preserve all of the interesting information.[9] The crucial point is that this procedure cannot exist in general, but only for certain classes of probability distributions. This is an introduction to the idea some kinds of structure in data are learnable from random examples, while other structures are not.

Consider the (Boltzmann) probability distribution for the states of a system in thermal equilibrium. If we expand the Hamiltonian as a sum of terms (operators) then the family of possible probability distributions is an exponential family in which the coupling constants for each operator are the parameters analogous to the λ_i above. In principle there could be an infinite number of these operators, but for a given class of systems we usually find that only a finite set are "relevant" in the renormalization group sense: if we write an effective Hamiltonian for coarse grained degrees of freedom, then only a finite number of terms will survive the coarse graining procedure. If we have only a finite number of terms in the Hamiltonian, then the family of Boltzmann distributions has sufficient statistics, which are just the expectation values of the relevant operators. This means that the expectation values of the relevant operators carry all the information that the (coarse grained) configuration of the system can provide about the coupling constants, which in turn is information about the identity or microscopic structure of the system. Thus the statement that there are only a finite number of relevant operators

[8] These conditions mean that the sufficient statistics can be constructed as running averages, and that the same data points in different sequence carry the same information.

[9] There are more subtle questions: How many bits are we squeezing out in this process of compression, and how many bits of relevance are left? We return to this and related questions when we talk about learning later in the course.

is also the statement that a finite number of expectation values carries all the information about the microscopic dynamics. The 'if' part of this statement is obvious: if there are only a finite number of relevant operators, then the expectation values of these operators carry all the information about the identity of the system. The statisticians, through the theorem about the uniqueness of exponential families, give us the 'only if': a finite number of expectation values (or correlation functions) can provide all the information about the system *only if* the effective Hamiltonian has a finite number of relevant operators. I suspect that there is more to say along these lines, but let us turn instead to some examples.

Consider the situation in which the data D are real numbers x. Suppose that we know the mean value of x and its variance. This is equivalent to knowledge of two expectation values,

$$\bar{f}_1 = \langle x \rangle = \int dx P(x) x, \quad \text{and} \tag{57}$$

$$\bar{f}_2 = \langle x^2 \rangle = \int dx P(x) x^2, \tag{58}$$

so we have $f_1(x) = x$ and $f_2(x) = x^2$. Thus, from Eq. (46), the maximum entropy distribution is of the form

$$P(x) = \frac{1}{Z} \exp(-\lambda_1 x - \lambda_2 x^2). \tag{59}$$

This is a funny way of writing a more familiar object. If we identify the parameters $\lambda_2 = 1/(2\sigma^2)$ and $\lambda_1 = -\langle x \rangle / \sigma^2$, then we can rewrite the maximum entropy distribution as the usual Gaussian,

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2} (x - \langle x \rangle)^2 \right]. \tag{60}$$

We recall that Gaussian distributions usually arise through the central limit theorem: if the random variables of interest can be thought of as sums of many independent events, then the distributions of the observable variables converge to Gaussians. This provides us with a 'mechanistic' or reductionist view of why Gaussians are so important. A very different view comes from information theory: if all we know about a variable is the mean and the variance, then the Gaussian distribution is the maximum entropy distribution consistent with this knowledge. Since the entropy measures (returning to our physical intuition) the randomness or disorder of the system, the Gaussian distribution describes the 'most random' or 'least structured' distribution that can generate the known mean and variance.

A somewhat different situation is when the data D are generated by counting. Then the relevant variable is an integer $n = 0, 1, 2, \cdots$, and it is natural to imagine that what we know is the mean count $\langle n \rangle$. One way this problem can arise is that we are trying to communicate and are restricted to sending discrete or quantized units. An obvious case is in optical communication, where the quanta are photons. In the brain, quantization abounds: most neurons do not generate continuous analog voltages but rather communicate with one another through stereotyped pulses or spikes, and even if the voltages vary continuously transmission across a synapse involves the release of a chemical transmitter which is packaged into discrete vesicles. It can be relatively easy to measure the mean rate at which discrete events are counted, and we might want to know what bounds this mean rate places on the ability of the cells to convey information. Alternatively, there is an energetic cost associated with these discrete events—generating the electrical currents that underlie the spike, constructing and filling the vesicles, ... —and we might want to characterize the mechanisms by their cost per bit rather than their cost per event. If we know the mean count, there is (as for the Boltzmann distribution) only one function $f_1(n) = n$ that can appear in the exponential of the distribution, so that

$$P(n) = \frac{1}{Z} \exp(-\lambda n). \tag{61}$$

Of course we have to choose the Lagrange multiplier to fix the mean count, and it turns out that $\lambda = \ln(1 + 1/\langle n \rangle)$; further we can find the entropy

$$S_{\max}(\text{counting}) = \log_2(1 + \langle n \rangle) + \langle n \rangle \log_2(1 + 1/\langle n \rangle). \tag{62}$$

The information conveyed by counting something can never exceed the entropy of the distribution of counts, and if we know the mean count then the entropy can never exceed the bound in Eq. (62). Thus, if we have a system in which information is conveyed by counting discrete events, the simple fact that we count only a limited number of events (on average) sets a bound on how much information can be transmitted. We will see that real neurons and synapses approach this fundamental limit.

One might suppose that if information is coded in the counting of discrete events, then each event carries a certain amount of information. In fact this is not quite right. In particular, if we count a large number of events then the maximum counting entropy becomes

$$S_{\max}(\text{counting}; \langle n \rangle \to \infty) \sim \log_2(\langle n \rangle e), \tag{63}$$

and so we are guaranteed that the entropy (and hence the information) per event goes to zero, although the approach is slow. On the other hand, if events are very rare, so that the mean count is much less than one, we find the maximum entropy per event

$$\frac{1}{\langle n \rangle} S_{\max}(\text{counting}; \langle n \rangle << 1) \sim \log_2 \left(\frac{e}{\langle n \rangle} \right), \tag{64}$$

which is arbitrarily large for small mean count. This makes sense: rare events have an arbitrarily large capacity to surprise us and hence to convey information. It is important to note, though, that the maximum entropy per event is a monotonically decreasing function of

the mean count. Thus if we are counting spikes from a neuron, counting in larger windows (hence larger mean counts) is always less efficient in terms of bits per spike.

If it is more efficient to count in small time windows, perhaps we should think not about counting but about measuring the arrival times of the discrete events. If we look at a total (large) time interval $0 < t < T$, then we will observe arrival times $t_1, t_2, \cdots, t_N$ in this interval; note that the number of events N is also a random variable. We want to find the distribution $P(t_1, t_2, \cdots, t_N)$ that maximizes the entropy while holding fixed the average event rate. We can write the entropy of the distribution as a sum of two terms, one from the entropy of the arrival times given the count and one from the entropy of the counting distribution:

$$\begin{aligned} S &\equiv -\sum_{N=0}^{\infty} \int d^N t_n P(t_1, t_2, \cdots, t_N) \log_2 P(t_1, t_2, \cdots, t_N) \\ &= \sum_{N=0}^{\infty} P(N) S_{\rm time}(N) - \sum_{N=0}^{\infty} P(N) \log_2 P(N), \end{aligned} \tag{65}$$

where we have made use of

$$P(t_1, t_2, \cdots, t_N) = P(t_1, t_2, \cdots, t_N | N) P(N), \tag{66}$$

and the (conditional) entropy of the arrival times in given by

$$S_{\rm time}(N) = -\int d^N t_n P(t_1, t_2, \cdots, t_N | N) \log_2 P(t_1, t_2, \cdots, t_N | N). \tag{67}$$

If all we fix is the mean count, $\langle N \rangle = \sum_N P(N) N$, then the conditional distributions for the locations of the events given the total number of events, $P(t_1, t_2, \cdots, t_N | N)$, are unconstrained. We can maximize the contribution of each of these terms to the entropy [the terms in the first sum of Eq. (65)] by making the distributions $P(t_1, t_2, \cdots, t_N | N)$ uniform, but it is important to be careful about normalization. When we integrate over all the times $t_1, t_2, \cdots, t_N$, we are forgetting that the events are all identical, and hence that permutations of the times describe the same events. Thus the normalization condition is *not*

$$\int_0^T dt_1 \int_0^T dt_2 \cdots \int_0^T dt_N P(t_1, t_2, \cdots, t_N | N) = 1, \tag{68}$$

but rather

$$\frac{1}{N!} \int_0^T dt_1 \int_0^T dt_2 \cdots \int_0^T dt_N P(t_1, t_2, \cdots, t_N | N) = 1. \tag{69}$$

This means that the uniform distribution must be

$$P(t_1, t_2, \cdots, t_N | N) = \frac{N!}{T^N}, \tag{70}$$

and hence that the entropy [substituting into Eq. (65)] becomes

$$S = -\sum_{N=0}^{\infty} P(N) \left[\log_2 \left(\frac{N!}{T^N} \right) + \log_2 P(N) \right]. \tag{71}$$

Now to find the maximum entropy we proceed as before. We introduce Lagrange multipliers to constrain the mean count and the normalization of the distribution $P(N)$, which leads to the function

$$\tilde{S} = -\sum_{N=0}^{\infty} P(N) \left[\log_2 \left(\frac{N!}{T^N} \right) + \log_2 P(N) + \lambda_0 + \lambda_1 N \right], \tag{72}$$

and then we maximize this function by varying $P(N)$. As before the different Ns are not coupled, so the optimization conditions are simple:

$$0 = \frac{\partial \tilde{S}}{\partial P(N)} \tag{73}$$

$$= -\frac{1}{\ln 2} \left[\ln \left(\frac{N!}{T^N} \right) + \ln P(N) + 1 \right] - \lambda_0 - \lambda_1 N, \tag{74}$$

$$\ln P(N) = -\ln\left(\frac{N!}{T^N}\right) - (\lambda_1 \ln 2)N - (1 + \lambda_0 \ln 2). \quad (75)$$

Combining terms and simplifying, we have

$$P(N) = \frac{1}{Z}\frac{(\lambda T)^N}{N!}, \quad (76)$$

$$Z = \sum_{N=0}^{\infty} \frac{(\lambda T)^N}{N!} = \exp(\lambda T). \quad (77)$$

This is the Poisson distribution.

The Poisson distribution usually is derived (as in our discussion of photon counting) by assuming that the probability of occurrence of an event in any small time bin of size $\Delta\tau$ is independent of events in any other bin, and then we let $\Delta\tau \to 0$ to obtain a distribution in the continuum. This is not surprising: we have found that the maximum entropy distribution of events given the mean number of events (or their density $\langle N\rangle/T$) is given by the Poisson distribution, which corresponds to the events being thrown down at random with some probability per unit time (again, $\langle N\rangle/T$) and no interactions among the events. This describes an 'ideal gas' of events along a line (time). More generally, the ideal gas is the gas with maximum entropy given its density; interactions among the gas molecules always reduce the entropy if we hold the density fixed.

If we have multiple variables, $x_1, x_2, \cdots, x_N$, then we can go through all of the same analyses as before. In particular, if these are continuous variables and we are told the means and covariances among the variables, then the maximum entropy distribution is again a Gaussian distribution, this time the appropriate multidimensional Gaussian. This example, like the other examples so far, is simple in that we can give not only the form of the distribution but we can find the values of the parameters that will satisfy the constraints. In general this is not so easy: think of the Boltzmann distribution, where we would have to adjust the temperature to obtain a given value of the average energy, but if we can give an explicit relation between the temperature and average energy for any system then we have solved almost all of statistical mechanics!

One important example is provided by binary strings. If we label 1s by spin up and 0s by spin down, the binary string is equivalent to an Ising chain $\{\sigma_i\}$. Fixing the probability of a 1 is the same as fixing the mean magnetization $\langle\sigma_i\rangle$. If, in addition, we specify the joint probability of two 1s occurring in bins separated by n steps (for all n), this is equivalent to fixing the spin–spin correlation function $\langle\sigma_i\sigma_j\rangle$. The maximum entropy distribution consistent with these constraints is an Ising model,

$$P[\{\sigma_i\}] = \frac{1}{Z}\exp\left[-h\sum_i \sigma_i - \sum_{ij} J_{ij}\sigma_i\sigma_j\right]; \quad (78)$$

note that the interactions are pairwise (because we fix only a two–point function) but not limited to near neighbors. Obviously the problem of finding the exchange interactions which match the correlation function is not so simple.

Another interesting feature of the Ising or binary string problem concerns higher order correlation functions. If we have continuous variables and constrain the two–point correlation functions, then the maximum entropy distribution is Gaussian and there are no nontrivial higher order correlations. But if the signals we observe are discrete, as in the sequence of spikes from a neuron, then the maximum entropy distribution is an Ising model and this model makes nontrivial predictions about the multipoint correlations. In particular, if we record the spike trains from K separate neurons and measure all of the pairwise correlation functions, then the corresponding Ising model predicts that there will be irreducible correlations among triplets of neurons, and higher order correlations as well [Schneidman et al 2006].

Before closing the discussion of maximum entropy distributions, note that our simple solution to the problem, Eq. (46), might not work. Taking derivatives and setting them to zero works only if the solution to our problem is in the interior of the domain allowed by the constraints. It is also possible that the solution lies at the boundary of this allowed region. This seems especially likely when we combine different kinds of constraints, such as trying to find the maximum entropy distribution of images consistent both with the two–point correlation function and with the histogram of intensity values at one point. The relevant distribution is a 2D field theory with a (generally nonlocal) quadratic 'kinetic energy' and some arbitrary local potential; it is not clear that all combinations of correlations and histograms can be realized, nor that the resulting field theory will be stable under renormalization.[10] There are many open questions here.

F. Information transmission with noise

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:

$$y = gx + \xi. \quad (79)$$

[10] The empirical histograms of local quantities in natural images *are* stable under renormalization [Ruderman and Bialek 1994].

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 ξ is Gaussian noise means that once we know x, y is Gaussian distributed around a mean value of gx:

$$P(y|x) = \frac{1}{\sqrt{2\pi\langle\xi^2\rangle}} \exp\left[-\frac{1}{2\langle\xi^2\rangle}(y-gx)^2\right]. \quad (80)$$

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

$$P(x) = \frac{1}{\sqrt{2\pi\langle x^2\rangle}} \exp\left[-\frac{1}{2\langle x^2\rangle}x^2\right], \quad (81)$$

and hence y itself is Gaussian,

$$P(y) = \frac{1}{\sqrt{2\pi\langle y^2\rangle}} \exp\left[-\frac{1}{2\langle y^2\rangle}y^2\right] \quad (82)$$

$$\langle y^2\rangle = g^2\langle x^2\rangle + \langle\xi^2\rangle. \quad (83)$$

To compute the information that y provides about x we use Eq. (26):

$$I(y\to x) = \int dy \int dx P(x,y)\log_2\left[\frac{P(x,y)}{P(x)P(y)}\right] \text{ bits} \quad (84)$$

$$= \frac{1}{\ln 2}\int dy \int dx P(x,y)\ln\left[\frac{P(y|x)}{P(y)}\right] \quad (85)$$

$$= \frac{1}{\ln 2}\left\langle \ln\left[\frac{\sqrt{2\pi\langle y^2\rangle}}{\sqrt{2\pi\langle\xi^2\rangle}}\right] - \frac{1}{2\langle\xi^2\rangle}(y-gx)^2 + \frac{1}{2\langle y^2\rangle}y^2\right\rangle, \quad (86)$$

where by $\langle\cdots\rangle$ we understand an expectation value over the joint distribution $P(x,y)$. Now in Eq. (86) 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,

$$I(y\to x) = \frac{1}{\ln 2}\left[\ln\sqrt{\frac{\langle y^2\rangle}{\langle\xi^2\rangle}} - \frac{1}{2} + \frac{1}{2}\right] \quad (87)$$

$$= \frac{1}{2}\log_2\left(\frac{\langle y^2\rangle}{\langle\xi^2\rangle}\right) \quad (88)$$

$$= \frac{1}{2}\log_2\left(1 + \frac{g^2\langle x^2\rangle}{\langle\xi^2\rangle}\right) \text{ bits.} \quad (89)$$

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:

$$y = g(x + \eta_{\rm eff}), \quad (90)$$

where, obviously, $\eta_{\rm eff} = \xi/g$. Note that the "effective noise" $\eta_{\rm eff}$ 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. Clearly 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 want to know the variance in the pulse rate; 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. With the effective noise level, the information transmission takes a simple form,

$$I(y\to x) = \frac{1}{2}\log_2\left(1 + \frac{\langle x^2\rangle}{\langle\eta_{\rm eff}^2\rangle}\right) \text{ bits,} \quad (91)$$

or

$$I(y\to x) = \frac{1}{2}\log_2(1+SNR), \quad (92)$$

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

The result in Eq. (92) 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 $\delta x_1 \sim \langle\eta^2_{\rm eff}\rangle^{1/2}$, 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.

As a next step consider the case where we observe several variables $y_1, y_2, \cdots, y_K$ in the hopes of learning about the same number of underlying signals $x_1, x_2, \cdots, x_K$. The equations analogous to Eq. (79) are then

$$y_{\rm i} = g_{\rm ij}x_{\rm j} + \xi_{\rm i}, \tag{93}$$

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

$$S_{\rm ij} = \langle x_{\rm i}x_{\rm j}\rangle \tag{94}$$

$$N_{\rm ij} = \langle \xi_{\rm i}\xi_{\rm j}\rangle. \tag{95}$$

The relevant probability distributions are

$$P(\{x_{\rm i}\}) = \frac{1}{\sqrt{(2\pi)^K \det S}}\exp\left[-\frac{1}{2}x_{\rm i}\cdot(S^{-1})_{\rm ij}\cdot x_{\rm j}\right] \tag{96}$$

$$P(\{y_{\rm i}\}|\{x_{\rm i}\}) = \frac{1}{\sqrt{(2\pi)^K \det N}}\exp\left[-\frac{1}{2}(y_{\rm j}-g_{\rm ik}x_{\rm k})\cdot(N^{-1})_{\rm ij}\cdot(y_{\rm j}-g_{\rm jm}x_{\rm m})\right], \tag{97}$$

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

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

$$P(\{y_{\rm i}\}) = \int d^K x P(\{y_{\rm i}\}|\{x_{\rm i}\})P(\{x_{\rm i}\}), \tag{98}$$

and then we write the information as an expectation value,

$$I(\{y_{\rm i}\}\to\{x_{\rm i}\}) = \left\langle \log_2\left[\frac{P(\{y_{\rm i}\}|\{x_{\rm i}\})}{P(\{y_{\rm i}\})}\right]\right\rangle, \tag{99}$$

where $\langle\cdots\rangle$ denotes an average over the joint distribution $P(\{y_{\rm i}\},\{x_{\rm i}\})$. As in Eq. (86), 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

$$I(\{y_{\rm i}\}\to\{x_{\rm i}\}) = \frac{1}{2}{\rm Tr}\log_2[\mathbf{1} + N^{-1}\cdot(g\cdot S\cdot g^T)], \tag{100}$$

where 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 ys to make this matrix diagonal, which corresponds to finding the eigenvectors and eigenvalues of the covariance matrix; these eigenvectors are also called "principal components." 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 Tr log is equivalent to computing the information transmission along each direction using Eq. (92), and then summing the results.

An important case is when the different variables $x_{\rm i}$ 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. (93) we have the continuous linear response of the detector to input signals,

$$y(t) = \int dt' M(t-t')x(t') + \xi(t). \tag{101}$$

In this continuous case the analog of the covariance matrix $\langle x_{\rm i}x_{\rm j}\rangle$ is the correlation function $\langle x(t)x(t')\rangle$. We are usually interested in signals (and noise) that are stationary. This means 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,

$$\langle x(t)x(t')\rangle = C_x(t-t'). \tag{102}$$

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. (101), we again refer noise to the input by writing

$$y(t) = \int dt' M(t-t')[x(t') + \eta_{\rm eff}(t')]. \quad (103)$$

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

$$SNR(\omega) = \frac{S_x(\omega)}{N_{\rm eff}(\omega)}, \quad (104)$$

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

$$I[y(t) \to x(t)] = \frac{1}{2}\sum_\omega \log_2[1 + SNR(\omega)]. \quad (105)$$

Finally, to compute the frequency sum, we recall that

$$\sum_{\rm n} f(\omega_{\rm n}) \to T\int \frac{d\omega}{2\pi} f(\omega). \quad (106)$$

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

$$I[y(0<t<T) \to x(0<t<T)] \to \frac{T}{2}\int \frac{d\omega}{2\pi}\log_2[1+SNR(\omega)] \text{ bits}. \quad (107)$$

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

$$R_{\rm info} \equiv \lim_{T\to\infty}\frac{1}{T}\cdot I[y(0<t<T)\to x(0<t<T)] \quad (108)$$

$$= \frac{1}{2}\int\frac{d\omega}{2\pi}\log_2[1+SNR(\omega)] \text{ bits/sec}. \quad (109)$$

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 τ_0 then the maximum (Nyquist) frequency is $|\omega|_{\rm max} = \pi/\tau_0$.

The Gaussian channel is interesting in part because we can work everything out in detail, but in fact we can learn a little bit more than this. There are many situations in which there are good physical reasons to believe that noise will be Gaussian—it arises from the superposition of many small events, and the central limit theorem applies. If we know that noise is Gaussian and we know its spectrum, we might ask how much information can be gained about a signal that has some known statistical properties. Because of the maximum entropy property of the Gaussian distribution, this true information transmission rate is always less than or equal to what we would compute by assuming that signal is Gaussian, measuring its spectrum, and plugging into Eq. (109). Notice that this bound is saturated only in the case where the signal in fact is Gaussian, that is when the signal has some of the same statistical structure as the noise. We will see another example of this somewhat counterintuitive principle in just a moment.

Now we can use the maximum entropy argument in a different way. When we study cells deep in the brain, we might choose to deliver signals that are drawn from a Gaussian distribution, but given the nonlinearities of real neurons there is no reason to think that the effective noise in the representation of these signals will be Gaussian. But we can use the maximum entropy property of the Gaussian once again, this time to show that if can measure the power spectrum of the effective noise, and we plug this into Eq. (109), then we will obtain a *lower* bound to the true information transmission rate. Thus we can make conservative statements about what neurons can do, and we will see that even these conservative statements can be quite powerful.

If the effective noise is Gaussian, then we know that the maximum information transmission is achieved by choosing signals that are also Gaussian. But this doesn't tell us how to choose the spectrum of these maximally informative signals. We would like to say that measurements of effective noise levels can be translated into bounds on information transmission, but this requires that we solve the optimization problem for shaping the spectrum. 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 informa-

tion 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 $R_{\rm info}$ while holding $\langle x^2\rangle$ fixed. As before, we introduce a Lagrange multiplier and maximize a new function

$$\tilde{R} = R_{\rm info} - \lambda\langle x^2\rangle \quad (110)$$

$$= \frac{1}{2}\int\frac{d\omega}{2\pi}\log_2\left[1+\frac{S_x(\omega)}{N_{\rm eff}(\omega)}\right] - \lambda\int\frac{d\omega}{2\pi}S_x(\omega). \quad (111)$$

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

$$\frac{\delta\tilde{R}}{\delta S_x(\omega)} = \frac{1}{2\ln 2}\cdot\frac{1}{1+S_x(\omega)/N_{\rm eff}(\omega)}\cdot\frac{1}{N_{\rm eff}(\omega)} - \lambda, \quad (112)$$

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

$$S_x(\omega) + N_{\rm eff}(\omega) = \frac{1}{2\lambda\ln 2}. \quad (113)$$

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. (113) 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. (113) 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.

G. Back to the fly's retina

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:

$$V(t) = V_{\rm DC} + \int dt' T(t-t')C(t') + \delta V(t). \quad (114)$$

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

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

$$V(t) = V_{\rm DC} + \int dt' T(t-t')[C(t') + \delta C_{\rm eff}(t)]. \quad (115)$$

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

$$\langle\delta C_{\rm eff}(t)\delta C_{\rm eff}(t')\rangle = \int\frac{d\omega}{2\pi}N_C^{\rm eff}(\omega)\exp[-i\omega(t-t')]. \quad (116)$$

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 $N_C^{\rm eff}(\omega)$, 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 evolution!). This is just the optimization discussed above, so it is possible to turn the measurements on signals and

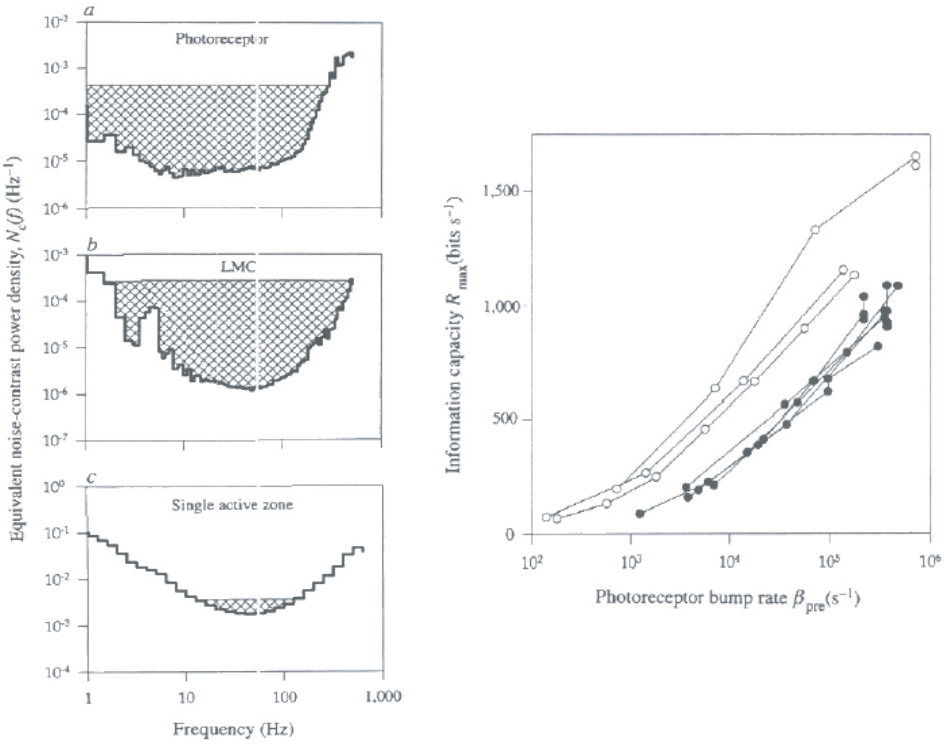

FIG. 2 At left, the effective contrast noise levels in a single photoreceptor cell, a single LMC (the second order cell) and the inferred noise level for a single active zone of the synapse from photoreceptor to LMC. The hatching shows the signal spectra required to whiten the total output over the largest 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].

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 2.

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 is 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, from Eq. (62) above. 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].

These two examples [Laughlin 1981; de Ruyter van Steveninck & Laughlin 1996], both from the first synapse in fly vision, provide 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 (e.g., why does the system choose the particular filter characteristics that it does?) and there are gaps (e.g., putting together the dynamics and the nonlinearities). But, armed with some suggestive results, let's go on ... .