IV. EFFICIENT REPRESENTATION

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. 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. In 1948, Shannon proved a theorem that makes the connection precise: 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 Chapter we'll explore the possibility that biological systems in fact form efficient representations, maximizing the amount of relevant information that they transmit and process, subject to fundamental physical constraints.

The idea that a mathematically precise notion of "information" would be useful in thinking about the representation of information in the brain came very quickly after Shannon's original work. There is, therefore, a well developed set of ideas about the how many bits are carried by the responses of neurons, in what sense the encoding of sensory signals into sequences of action potentials is efficient, and so on. More subtly, there is a body of work on the theory of learning that can be summarized by saying that the goal of learning is to build an efficient representation of what we have seen. In contrast, most discussions of signaling and control at the molecular level has left "information" as a colloquial concept. One of the goals of this Chapter, then, is to bridge this gap. Hopefully, in the physics tradition, it will be clear how the same concepts can be used in thinking about the broadest possible range of phenomena. We begin, however, with the foundations.

A. 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. Let us start 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. Thus, if the probability of the n^{th} possible response is p_n, then the entropy is

$$S = -\sum_n p_n \log_2 p_n \text{ bits}. \tag{616}$$

Our intuition from statistical mechanics suggests that the entropy S measures Max's uncertainty about what Allan will say in response to his question, in the same way that the entropy of a gas measures our lack of knowledge about the microstates of all the constituent molecules. Once Allan gives his answer, all of 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.

If we want to have a general measure of how much information is gained on hearing the answer to a question, we have to put aside the details of the questions and the answers—although this might make us uncomfortable,

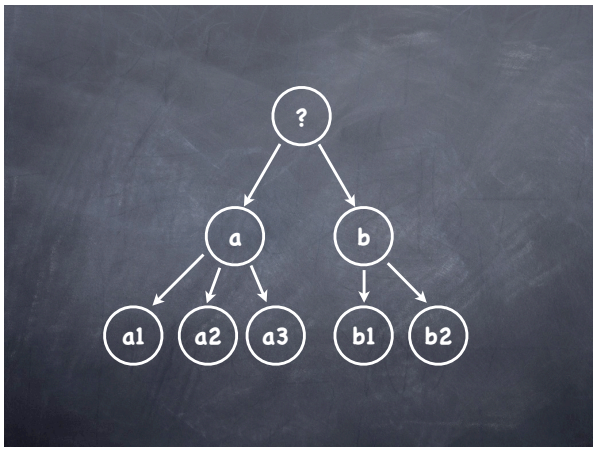

FIG. 128 The branching postulate in Shannon's proof. The idea is to break a big question into multiple parts, as in the familiar game of twenty questions. We start with some initial question, at the top (?). Depending on the answer to this question (a or b), we ask a new question. This second question in turn has multiple possible answers ($a1, a2, a3$ or $b1, b2$). In this tree structure, the various sub–questions live at branch points, with the answers emerging along the branches; finding our way to the full answer means following one path through the tree. The average information that we gain along this path should be additive, the weighted sum of information gained at every branch point.

and is something we should revisit. If we leave out the text of the questions and answers themselves, then all that remains are the probabilities $p_{\rm n}$ of hearing the different answers, and so Shannon assumes that the information gained must be a function of these probabilities, $I(\{p_{\rm n}\})$. The challenge is to determine this function.[76]

The first constraint is that, if all N possible answers are equally likely, then the information gained should be a monotonically increasing function of N—we learn more by asking questions that have a wider range of possible answers. 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, as in Fig 128, 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 $\{p_{\rm n}\}$ consistent with these three postulates—monotonicity, independence, and branching—is the entropy S, up to a multiplicative constant. The proof is sufficiently simple that it seems worth going through the details, not least to be sure we understand how little is required to derive such a powerful result.

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 $I(\{p_{\rm n}\}) = f(N)$. 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. But we have assumed that information from independent questions and answers must add, so the function $f(N)$ must obey the condition

$$f(k^m) = mf(k). \tag{617}$$

Notice that although we are focusing on cases where $N = k^m$, we have a condition that involves $f(k)$ for arbitrary k. It is easy to see that $f(N) \propto \log N$ satisfies this equation. To show that this is the unique solution, consider another pair of integers ℓ and n such that

$$k^m \le \ell^n \le k^{m+1}, \tag{618}$$

or, taking logarithms,

$$\frac{m}{n} \le \frac{\log \ell}{\log k} \le \frac{m}{n} + \frac{1}{n}. \tag{619}$$

Now because the information measure $f(N)$ is monotonically increasing with N, the ordering in Eq. (618) means that

$$f(k^m) \le f(\ell^n) \le f(k^{m+1}), \tag{620}$$

and hence from Eq. (617) we obtain

$$mf(k) \le nf(\ell) \le (m+1)f(k). \tag{621}$$

Dividing through by $nf(k)$ we have

$$\frac{m}{n} \le \frac{f(\ell)}{f(k)} \le \frac{m}{n} + \frac{1}{n}, \tag{622}$$

which is very similar to Eq. (619). 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. (619) is telling us that

$$\left| \frac{m}{n} - \frac{\log \ell}{\log k} \right| < \epsilon, \tag{623}$$

and Eq. (622) for the function $f(N)$ can similarly be rewritten as

$$\left| \frac{m}{n} - \frac{f(\ell)}{f(k)} \right| < \epsilon. \tag{624}$$

Putting these together, we have

$$\left| \frac{f(\ell)}{f(k)} - \frac{\log \ell}{\log k} \right| \le 2\epsilon, \tag{625}$$

so that $f(N) \propto \log N$ as promised. Note that if we were allowed to consider $f(N)$ as a continuous function, then we could have made a much simpler argument. But, strictly speaking, $f(N)$ is defined only at integer arguments.

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 in statistical mechanics: what are the units of entropy? In a 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.[77] In physics

[76] Notice that Shannon's 'zeroth' assumption—that the information gained is a function of the probability distribution over the answers to our question—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 literally unimaginable answer to our question. It is interesting to think about whether this is a real restriction.

[77] I have to admit that 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.

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 measuring information 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. Consider first the situation where all the probabilities are rational, that is

$$p_{\rm n} = \frac{k_{\rm n}}{\sum_{\rm m} k_{\rm m}}, \tag{626}$$

where all the $k_{\rm n}$ are integers. 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, as in Fig 129. To specify the full answer, we would first tell which group it is in, then tell which of the $k_{\rm n}$ possibilities is realized. In this two step process, at the first step we get the information we

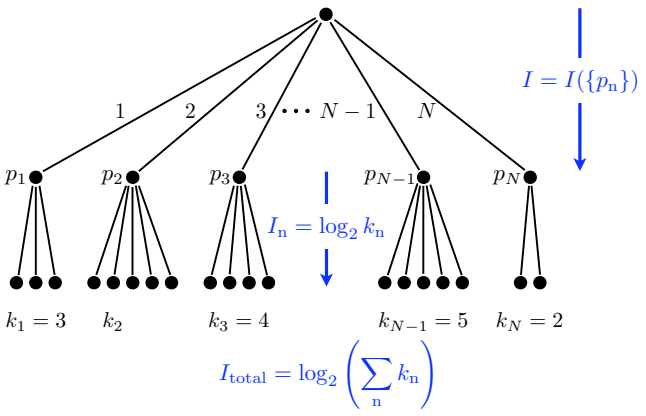

FIG. 129 Grouping. To determine the information gained with unequal probabilities, we consider a "big question" with answer that fall into N groups. By hypothesis, in each the $k_{\rm n}$ answers are equally likely.

are really looking for—which of the N groups are we in—and so the information in the first step is our unknown function,

$$I_1 = I(\{p_{\rm n}\}). \tag{627}$$

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

$$I_2 = \sum_{\rm n} p_{\rm n} I_{\rm n} = \sum_{\rm n} p_{\rm n} \log_2 k_{\rm n}. \tag{628}$$

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

$$I_1 + I_2 = \log_2 \left(N_{\rm total} \right), \tag{629}$$

$$I(\{p_{\rm n}\}) + \sum_{\rm n} p_{\rm n} \log_2 k_{\rm n} = \log_2 \left(\sum_{\rm m} k_{\rm m} \right). \tag{630}$$

Rearranging the terms, we find

$$I(\{p_{\rm n}\}) = -\sum_{\rm n} p_{\rm n} \log_2 \left(\frac{k_{\rm n}}{\sum_{\rm m} k_{\rm m}} \right) \tag{631}$$

$$= -\sum_{\rm n} p_{\rm n} \log_2 p_{\rm n}. \tag{632}$$

Again, although this is worked out explicitly for the case where the $p_{\rm n}$ are rational, it must be the general answer if the information measure is continuous. So we are done: the average information gained on hearing the answer to a question is measured uniquely by the entropy of the distribution of possible answers.

It is worth pausing here to note that what Shannon did is very different from our conventional experience in using mathematics to describe the natural world. In most of physics, we have some set of observations (the motion of the planets in the night sky, for example) that can be made quantitative (as Brahe did), and we search for mathematical structures that can explain and unify these data (Kepler, Newton). In contrast, Shannon considered an everyday phenomenon for which we have a colloquial language, and asked if this language itself could be made mathematically precise, without reference to quantitative data. It is remarkable that this actually worked, and that Shannon's construction has, as we will see, so many consequences.

When we try to quantify the information we gain from hearing the answer to a question, it seems 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

$$S_{\rm continuum} = -\int dx P(x) \log_2 P(x), \qquad (633)$$

or some multidimensional generalization. The difficulty is that probability distributions for continuous variables have units—$P(x)$ 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.

Problem 127: Dimensionality and the scaling of the entropy. As written, Eq (633) doesn't really make sense, because we are taking the log of something with units. Suppose we try to clean this up, and make bins along the x axis, each bin of width Δx and the $n^{\rm th}$ bin centered at $x_{\rm n}$. Then if the bins are reasonably small, the probability of falling in the $n^{\rm th}$ bin is $p_{\rm n} = P(x_{\rm n})\Delta x$.

(a.) Show that if you calculate the entropy in the usual way, you find

$$S = -\sum_{\rm n} p_{\rm n} \log_2 p_{\rm n} = S_{\rm continuum} - \log_2(\Delta x) \qquad (634)$$

in the limit $\Delta x \to 0$. More generally, show that in D dimensions

$$S = -\sum_{\rm n} p_{\rm n} \log_2 p_{\rm n} = S_{\rm continuum} - D \log_2(\Delta x). \qquad (635)$$

The result in Eq (635) suggests that the scaling of the entropy with bin size provides a measure of the dimensionality D of the underlying space. This is especially interesting if the intrinsic dimensionality is different from the dimensionality we happen to be using in describing the system. As an example, if we describe a system by its position in a two dimensional space (x, y), but really the points fall on a curve, then the right answer is that the system is one dimensional, not two dimensional.

(b.) Write a small program in MATLAB to generate 10^6 points in the (x, y) plane that fall on the circle $x^2 + y^2 = 1$. Then divide the plane (you can confine your attention to the region $-2 < x < 2$, and similarly for y) into boxes of size $(\Delta x) \times (\Delta x)$, and estimate the fraction of points that fall in each box. From this estimate, compute the entropy, and see how it varies as a function of Δx. Can you identify the signature of the reduced dimensionality?

(c.) Suppose that you take the 10^6 points from (b) and add, to each point, a bit of noise in the x and y directions, for example Gaussian noise with a standard deviation of $\sigma = 0.05$. Repeat the calculation of the entropy vs. box size. If you look closely enough ($\Delta x \ll \sigma$) the underlying probability distribution really is two dimensional, since there is independent noise along x and y. But if your resolution is more coarse ($\Delta x \gg \sigma$) you won't be able to "see" the noise and the points will appear to fall on a circle, corresponding to a one dimensional distribution. Can you see this transition in the plot of $S(\Delta x)$?

The problem of defining the entropy for continuous variables is familiar in statistical mechanics.[78] 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 start with a classical formulation (as did Boltzmann and Gibbs, of course) then we would find ourselves with the problems of Eq. (633), 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.

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 we 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 we have to do in answering the question is to provide a unique label. If we think of the label as a 'code word' for the answer, then in this simple case the length of the code word that represents the $n^{\rm th}$ possible answer is given by $\ell_{\rm n} = -\log_2 p_{\rm n}$, and the average length of a code word is given by the entropy.

The equality of the entropy and the average length of code words is much more general than our simple example. Before proceeding, however, it is important to realize that the entropy is emerging as the answer to two

---

[78] Indeed, this problem is so troublesome that it has led to a serious shift in our teaching. It is simpler to define everything in the case where states are discrete, and this has led many people to argue that we shouldn't teach statistical physics until after students have learned quantum mechanics. Whatever advantages this might have, it guarantees that many US students never see anything statistical (beyond a few lectures on the kinetic theory of gases) until their third year of university, which is quite late.

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 code words of length $\ell_{\rm n} = -\log_2 p_{\rm n}$ provided that we can find a suitable 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. Let us label the sequences of answers $\mathrm{n}_1\mathrm{n}_2\cdots\mathrm{n}_N$, and these sequences have probabilites $P(\mathrm{n}_1\mathrm{n}_2\cdots\mathrm{n}_N)$.[79] From these probabilities we can compute an entropy that must depend on the length of the sequence,

$$S(N) = -\sum_{\mathrm{n}_1}\sum_{\mathrm{n}_2}\cdots\sum_{\mathrm{n}_N} P(\mathrm{n}_1\mathrm{n}_2\cdots\mathrm{n}_N)\log_2 P(\mathrm{n}_1\mathrm{n}_2\cdots\mathrm{n}_N). \tag{636}$$

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

$$\lim_{N\to\infty}\frac{S(N)}{N} = \mathcal{S}, \tag{637}$$

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.

Problem 128: Probabilities and the equivalence of ensembles.[80] Consider an ideal monatomic gas in three dimensions, for which the energy is

$$E = \frac{1}{2m}\sum_{\mathrm{i}=1}^{3N} p_{\mathrm{i}}^2, \tag{638}$$

where m is the atomic mass. We will define the classical sum over states to be an integral over positions and velocities, normalized by appropriate powers of Planck's constant h.

(a.) The partition function in the microcanonical ensemble is

$$Z_{\rm micro}(E) \equiv \frac{1}{h^{3N}}\int d^3x \int d^3p\,\delta\left(E - \frac{1}{2m}\sum_{\mathrm{i}=1}^{3N} p_{\mathrm{i}}^2\right) \tag{639}$$

$$= \left(\frac{V}{h^3}\right)^N \int d^3p\,\delta\left(E - \frac{1}{2m}\sum_{\mathrm{i}=1}^{3N} p_{\mathrm{i}}^2\right). \tag{640}$$

If the energy is fixed with precision ϵ, then $Z_{\rm micro}(E)\epsilon$ is the number of accessible states, all occurring with equal probability, and so the microcanonical entropy is $S_{\rm micro}(E) = \log_2[Z_{\rm micro}(E)\epsilon]$. Use the Fourier representation of the delta function and the method of

[79] Notice that, at this point, we do not need to assume that successive questions have independent answers.

[80] This should be a review of things you learned in a statistical mechanics class, though perhaps in slightly different language. It is useful to make all of this explicit here.

steepest descent to derive the asymptotic behavior of $S_{\rm micro}(E)$ at large N.

(b.) In the canonical ensemble, at inverse temperature β, the probability of being in any state is given by the Boltzmann distribution,

$$P = \frac{1}{Z(\beta)} e^{-\beta E}, \tag{641}$$

where

$$Z(\beta) = \frac{1}{h^{3N}} \int d^3x \int d^3p \, \exp\left(-\frac{\beta}{2m} \sum_{\rm i=1}^{3N} p_{\rm i}^2 \right). \tag{642}$$

Evaluate $Z(\beta)$ and the entropy $S(\beta)$. Review what we mean when we say that the entropy is the same in the canonical and microcanonical ensembles at large N.

(c.) The typical probability of a state in the canonical ensemble is $P_{\rm typical} = 2^{-S(\beta)}$. Define the deviation from this typical probability as $\Delta = \log_2(P/P_{\rm typical})$. What can you say about the distribution of Δ over all the states? Can you make a precise version of the statement that "most" states have either "almost" the typical probability or zero probability? For example, can you put a bound on the fraction f of states which have $|\Delta| > \delta_c$? How does the relation between f and δ_c change with N?

Problem 129: More about typicality. Consider drawing N samples of a variable that can take on K different values, with probabilities $p_1, p_2, \cdots, p_K$. Let the sequence of samples that you observe be called ${\rm i}_1, {\rm i}_2, \cdots, {\rm i}_N$, which has probability

$$P = \prod_{\rm n=1}^{N} p_{\rm i_n}. \tag{643}$$

It should be easy to show that the average of $L = -(1/N)\log_2 P$ is the entropy of the underlying distribution, $S = -\sum_{\rm i} p_{\rm i} \log_2 p_{\rm i}$. Say as much as you can about the distribution of L as N 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. An important if obvious consequence is that if we have many rather unlikely 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 $\ell_{\rm n} \approx -\log_2 p_{\rm n}$ bits to represent each individual answer n. Thus, even individual answers which are more surprising require more space to write down.

As a simple example, imagine that we have four answers, with probabilities $p_1 = 1/2$, $p_2 = 1/4$, and $p_3 = p_4 = 1/8$. Naively, if we use a binary representation we will need two bits to represent the four possibilities. But the entropy is

$$S \equiv \sum_{\rm i=1}^{4} p_{\rm i} \log_2 p_{\rm i} = \frac{1}{2}\log_2 2 + \frac{1}{4}\log_2 4 + \frac{2}{8}\log_2 8 = \frac{7}{4}, \tag{644}$$

which is less than two bits (as it must be). Suppose that we represent the four possibilities by the binary sequences:

$$1 \to 0, \tag{645}$$
$$2 \to 10, \tag{646}$$
$$3 \to 110, \tag{647}$$
$$4 \to 111. \tag{648}$$

Notice that the length of each code word obeys $\ell_{\rm i} = -\log_2 p_{\rm i}$, so we know that, on average, the number of binary digits that we use per answer will be equal to the entropy. This illustrates the idea that, by using code words of different lengths, we can reduce the average amount of space we need to write things down.

Problem 130: Do we need commas? When we represent a sequence of answers, we have to be sure that we can find the boundaries between the code words. If all the words have the same length, we can just count, but this doesn't work if we use unequal lengths. At worst, we could add an extra symbol to "punctuate" the stream of words, but this takes extra space and surely is inefficient. Convince yourself that the code defined by Eqs (645) through (648) does not need any extra symbols—all sequences of code words can be parsed uniquely.

To complete the picture, we have to put together the ideas of typicality and code words of varying length. Suppose that we look at a block of N answers, ${\rm n}_1, {\rm n}_2, \cdots, {\rm n}_N$ as before; let's label this block (or "state," to reinforce the analogy with statistical physics) by s, which occurs with probability p_s. We choose the labels so that all the states are numbered in order of their probability, that is $p_1 \geq p_2 \geq \cdots \geq p_K$, where K is the number of possible sequences of length N. For each state s we can compute the cumulative probability of lower energy (higher probability) states, $P_s \equiv \sum_{i=1}^{s-1} p_i$. Now take this cumulative probability and expand it as a binary number. If we stop after m_s digits, where

$$-\log_2 p_s \leq m_s < -\log_2 p_s + 1, \tag{649}$$

then we guarantee that this binary number we are looking at will be different from any subsequent number with larger s, so it is a unique encoding of the state s, as shown schematically in Fig 130. But now we can see that the average number of binary digits we have used to encode the blocks of length N will be

$$L(N) \equiv \sum_s p_s m_s, \tag{650}$$

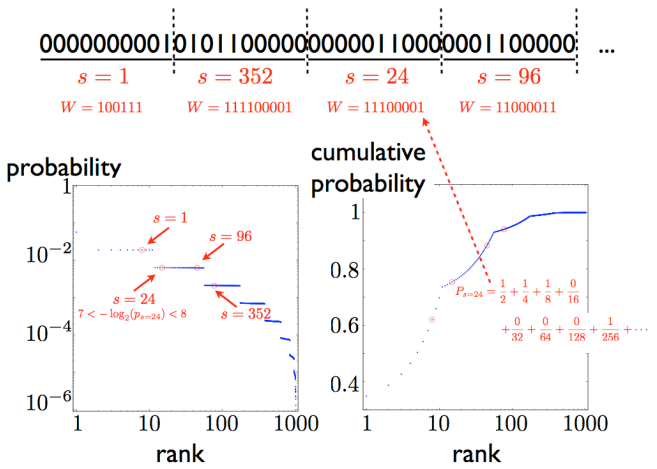

FIG. 130 Coding of sequences with variable word length. In a stream where '0' and '1' occur independently, but with unequal probabilities, we can compress our description by coding N–bit blocks; here $N = 10$. Each block can be labelled by s, the number equivalent to the binary string (top). These states have widely varying probability p_s (lower left). We can compute the cumulative probability of states with higher probability (lower right), as described in the text, and use the binary expansion of this cumulative probability as the code word W. We stop the expansion at a number of digits given by rounding up from $-\log_2(p_s)$.

and we can bound this from both sides,

$$\sum_s p_s m_s \left(-\log_2 p_s\right) \le L(N) < \sum_s p_s \left(-\log_2 p_s + 1\right) \quad (651)$$

$$S(N) \le L(N) < S(N) + 1, \quad (652)$$

where S is the entropy of the N–answer blocks, $S(N) = -\sum_s p_s \log_2 p_s$. If we count the length of the code *per answer*, then

$$\frac{S(N)}{N} \le \frac{L(N)}{N} < \frac{S(N)}{N} + \frac{1}{N}. \quad (653)$$

But, as before, we know that the entropy per degree of freedom should approach a finite entropy density, as in Eq (637), and now we see that the average code length per answer is within $1/N$ of this entropy density. Thus, as $N \to \infty$, the entropy and the minimum code length are equal.

To summarize, if we need to write down answers many times, then the minimum space required to write down these answers is, per answer, the entropy of the distribution out of which the answers are drawn. Notice that our choice of alphabet in which to write is arbitrary, but we also had an arbitrariness in choosing the units of entropy; this is the same arbitrariness. Thus, the statement that entropy is both the amount of information we gain and the amount of space we need to write down what we have learned is *not* arbitrary, and there are no constants floating around to spoil the exact equality. To reach this maximally compact representation, we must at least implicitly use the structure of the probability distribution out of which the answers are drawn, adjusting the lengths of individual code words in relation to the probability of the answer.

Problem 131: Coding rare events. Suppose that we have two possible answers, A and B, which occur with very unequal probabilities, $p_A \ll p_B$. Show that the entropy of the distribution of answers is approximately $S \approx p_A \log_2(e/p_A)$. If we have a long sequence of answers, most are B with a sprinkling of As. Try to encode such a sequence in binary form, using a code in which some symbol (e.g., 1111) is reserved for A, and the blocks of B are encoded by writing the number of consecutive Bs as a binary number. To make this work—that is, to be sure that your encoding can be uniquely decoded—you obviously have to be careful in the special case where the number of Bs is equal to 15 (1111 in binary form). Are there any other problems? Can you find a solution? Does this code come close to the lower bound on code length set by the entropy?

The idea that there is a minimum amount of space required to write down a description of a system is incredibly important. At a practical level, we pay for the resources needed to write things down, or to transmit information from one place to another, and so there is a premium on using as little space as possible. This is often called "data compression." More generally, this is the first indication that there is a general notion of efficiency in representing data, and we will see how this becomes relevant to biological systems.

The argument we have just given tells us that once we know the probability distribution for the states s, we have a code that we can use to represent these states, and asymptotically this code is of minimum length. Suppose that states really are chosen out of a distribution $\mathbf{p} \equiv \{p_s\}$, but we don't know this; instead, we think that the distribution is $\mathbf{q}$. Then (neglecting terms that are unimportant in the large N limit), we assign a code word of length $\ell_s = -\log_2 q_s$ to each state, and so the mean code length is

$$L = -\sum_s p_s \log_2 q_s. \quad (654)$$

This is different than the entropy of the distribution $\mathbf{p}$, and the difference

$$L - L_{\min} = L - S = -\sum_s p_s \log_2 q_s - \left[-\sum_s p_s \log_2 p_s\right] = \sum_s p_s \log_2 \left(\frac{p_s}{q_s}\right). \tag{655}$$

This quantity is zero if the two distributions are the same, and is positive for any pair of distributions $\mathbf{p}$ and $\mathbf{q}$; it is called the Kullback–Leibler divergence between the two distributions, and usually is written as

$$D_{KL}(\mathbf{p}||\mathbf{q}) = \sum_s p_s \log_2 \left(\frac{p_s}{q_s}\right). \tag{656}$$

Notice that this is not a symmetric quantity, and hence is not a metric on the space of distributions, although it does say something about the degree of similarity or difference between $\mathbf{p}$ and $\mathbf{q}$. D_{KL} also is sometimes called the "relative entropy" of the distribution $\mathbf{p}$ with respect to $\mathbf{q}$.

To emphasize the role of D_{KL} as a measure of difference between distributions, suppose that we are given N samples and have to decide whether they came from $\mathbf{p}$ or $\mathbf{q}$. Out of the N samples, n_1 come from state 1, n_2 come from state 2, and so on. So the probability that the distribution $\mathbf{p}$ generated these samples is

$$P(\text{samples}|\mathbf{p}) = A \prod_s p_s^{n_s}, \tag{657}$$

where A is a combinatorial factor, and similarly

$$P(\text{samples}|\mathbf{q}) = A \prod_s q_s^{n_s}. \tag{658}$$

What we want to know is, given the samples, what is the probability P that they came from the distribution $\mathbf{p}$ as opposed to $\mathbf{q}$? Let us say that, a priori, the two possibilities are equally likely. Then, by Bayes' rule,

$$P = \frac{P(\text{samples}|\mathbf{p})P(\mathbf{p})}{P(\text{samples})} \tag{659}$$

$$= \frac{P(\text{samples}|\mathbf{p})}{P(\text{samples}|\mathbf{p}) + P(\text{samples}|\mathbf{q})} \tag{660}$$

$$= \frac{1}{1+2^{-\Lambda}}, \tag{661}$$

where

$$\Lambda = \log_2 \left[\frac{P(\text{samples}|\mathbf{p})}{P(\text{samples}|\mathbf{q})}\right] = \sum_s n_s \log_2 \left(\frac{p_s}{q_s}\right). \tag{662}$$

As discussed in Chapter 1 [give specific pointer], Λ is called the log likelihood ratio. We notice that since it is proportional to all the n_s, it must also be proportional to N, and hence grows (on average) linearly with the number of samples. We can think of this as the accumulation of evidence for $\mathbf{p}$ vs. $\mathbf{q}$, and the rate at which this evidence accumulates is, asymptotically,

$$\lim_{N\to\infty} \frac{1}{N}\Lambda = \sum_s \left[\lim_{N\to\infty} \frac{n_s}{N}\right] \log_2 \left(\frac{p_s}{q_s}\right) \tag{663}$$

$$= \sum_s p_s \log_2 \left(\frac{p_s}{q_s}\right) \tag{664}$$

$$= D_{KL}(\mathbf{p}||\mathbf{q}). \tag{665}$$

Thus, the Kullback–Leibler divergence is, like the entropy itself, the answer to two very different questions: the cost of coding data using codes based on the wrong distribution, and the ease of discriminating the distributions from one another based on samples.

Problem 132: A little more about the Kullback–Leibler divergence.

(a.) Show that $D_{KL}(\mathbf{p}||\mathbf{q})$ is positive (semi–)definite, and is minimized when $\mathbf{p} = \mathbf{q}$.

(b.) $D_{KL}(\mathbf{p}||\mathbf{q})$ is unbounded, so some probability distributions are infinitely different from one another. Explain, using the connection to the accumulation of evidence, how to make sense out of this divergence.

(c.) If we have a family of distributions that depend on a parameter, $\mathbf{p}_\theta$, show that $D_{KL}(\mathbf{p}_\theta||\mathbf{p}_{\theta'})$ behaves as $F(\theta) \times (\theta - \theta')^2$ when the parameters θ and θ' are close. Give an explicit formula for $F(\theta)$.

(d.) Imagine that we draw N samples out of the distribution p_{θ_0}, but all we know is that the distribution is in the family p_θ. Use Bayes' rule to construct $P(\theta|\text{samples})$, and show that as N becomes large this becomes peaked around the right answer, $\theta = \theta_0$. Show that the variance around this peak is related to $F(\theta_0)$.

(e.) If the two distributions $\mathbf{p}$ and $\mathbf{q}$ are Gaussians, it's relatively easy to evaluate $D_{KL}(\mathbf{p}||\mathbf{q})$. Suppose that the two Gaussian differ in either their means or their variances, but not both. You should find that the choice of changing mean vs. variance makes a difference to the (a)symmetry of D_{KL}. Make this explicit, and use what we have shown about D_{KL} as a measure of discrimination to explain the origin of this difference.

The connection between entropy and information has (at least) one more very important consequence: correlations or order reduce the capacity to transmit information. Perhaps the most familiar example is in spelling. If all possible combinations of letters were legal words, then there would be $(26)^4 = 456,976$ four letter words. But if you look through a large, reasonably coherent body of English text—the collected works of a prolific author, or the last year of newspaper articles—you will find that there at most a few hundred four letter words being used. Most of this restriction of vocabulary comes from correlations among the letters in the word: once we have put

a 't' in the first position, it is much more likely that we will put a vowel in the second position; if we want to put a consonant then it has a high probability of being an 'h', and so on. It is important that, while correlations have signs—we speak both of correlation and anti–correlation—with respect to the entropy all correlations have the same effect, namely reducing the entropy. Indeed, as explained in [pointer to appendix on maximum entropy], we can construct models for the probability distribution of the states in a system that are consistent with some measured correlations but otherwise have the maximum possible entropy, and we can build a hierarchy of these models with ever smaller entropies as we take account of more correlations; once we capture all the relevant correlations, the entropy converges to its true value.

For four letter words, as an example, the entropy for random letters would be $S_{\rm rand} = 4\log_2(26) = 18.8$ bits. In the collected works of Jane Austen, the "one body" correlations, which measure unequal frequencies with which letters are used, reduces this to $S_{\rm ind} = 14$ bits. Taking account of the "two body" correlations between pairs of letters cuts this entropy nearly in half, to $S_2 = 7.48$ bits, while the true entropy of the distribution of four letter words in these texts in only slightly less, at $S = 6.92$ bits. Thus the entropy is nearly reduced by a factor of three from the case of completely random letters, and most of this reduction is explained by one and two body correlations. Again, the important point is that these correlations, which may have many advantages, certainly have the consequence of reducing our vocabulary and hence our capacity to transmit information.

This seems an appropriate moment to recall that entropy is a very old idea. It arises in thermodynamics first as a way of keeping track of heat flows, so that a small amount of heat dQ transferred at absolute temperature T generates a change in entropy $dS = dQ/T$. While there is no function Q which measure the heat content of a system, there is a function S that characterizes the (macroscopic) state of a system independent of the path to that state. But now we know that the entropy of a probability distribution also measures the amount of space that we need to write down a description of the (microscopic) states drawn out of that distribution.

[Would a schematic help here?] Let us imagine, then, a thought experiment in which we measure (with some fixed resolution) the positions and velocities of all the gas molecules in a small box, and type these numbers into a file on our computer. There are relatively efficient programs (gzip, or "compress" on a UNIX machine) that compresses such files to nearly their shortest possible length. If this really works, then the length of the file tells us the entropy of the distribution out of which the numbers in the file are being drawn, but this is the entropy of the gas. Thus, if we heat up the room by ten degrees, and repeat the process, we will find that the resulting data file is longer. More profoundly, if we measure the increase in the length of the file, we know the entropy change of the gas and hence the amount of heat that we had to add to the room in order to increase the temperature. This connection between a rather abstract quantity such as the length, in bits, of a computer file and a very tangible physical quantity such as the amount of heat added to a room has long struck me as one of the more dramatic, if elementary, examples of the power of mathematics to unify our description of very disparate phenomena.

Problem 133: Heat flows and file sizes. Give a problem that expands the thought experiment in the previous paragraph ... maybe with a polymer and entropic forces, where we can simulate?

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 $S[P(w)]$. 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 $S[P(w|d)]$ that is smaller than $S[P(w)]$ 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,

$$I(d \to w) \equiv S[P(w)] - S[P(w|d)]. \tag{666}$$

Notice that this depends on exactly what datum d we have observed.

Before proceeding, I should draw attention to some notational issues. Strictly speaking, entropy is a property of the probability distribution out of which the states of a system are drawn. Thus, we write $S[P(w)]$ to mean the entropy of the states of the world when these are drawn out of $P(w)$. Similarly, we should write $S[P(w|d)]$ for the entropy of states of the world conditional on having observed the data d. Notice that $S[\cdots]$ is the same functional in both cases. But, this is slightly cumbersome. Indeed, in statistical mechanics and thermodynamics we seldom talk about "the entropy of the distribution out of which the states of the gas have been drawn" (although we should); instead we just say "the entropy of the gas." In this spirit, sometimes I will write in the shorthand

$S(w) \equiv S[P(w)]$, and $S(w|d) \equiv S[P(w|d)]$. I hope this doesn't cause any confusion.

There is one more notational difficulty. When we talk about the states w of the world, it is natural to say that these states are drawn from the distribution $P(w)$. Similarly, when we talk about the data that we will collect, it is natural to write that particular observations d are drawn from the distribution $P(d)$. The problem is that $P(\cdot)$ refers to different functions in these two cases. We could solve this by noting carefully that the states of the world w come from a set of possible states, $w \in W$, and the distribution over these states should be written $P_W(w)$. Similarly, individual observations come from a set of possible observations, $d \in D$, and the distribution of these data should be written $P_D(d)$. Whenever there is a possibility for confusion, I'll try to adhere to this convention. In other cases, I'll slide to the more informal $P(w)$ and $P(d)$. Again, I hope this doesn't cause problems. [I am not sure that the current draft lives up to this policy, so please read carefully!]

With the notational issues settled, let's go back to our problem. Having defined the information gained in Eq (**??**), we should appreciate that this is not guaranteed to be positive. 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 $P(w)$, so that I_d will be negative. On the other hand, we 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 $P_D(d)$—we obtain the average information that d provides about w:

$$\langle (d \to w) \rangle = S(w) - \sum_d P_D(d) S(w|D). \tag{667}$$

This can be rearranged and simplified, and the result is so important that it is worth being very explicit about the algebra:

$$\begin{aligned}
\langle (d \to w) \rangle &= -\sum_w P_W(w) \log_2 P_W(w) - \sum_d P_D(d) \left[-\sum_w P(w|d) \log_2 P(w|d) \right] && (668)\\
&= -\sum_w \sum_d P(w, D) \log_2 P_W(w) + \sum_w \sum_d P(w|D) P_D(d) \log_2 P(w|d) && (669)\\
&= -\sum_w \sum_d P(w, D) \log_2 P_W(w) + \sum_w \sum_d P(w, D) \log_2 P(w|d) && (670)\\
&= \sum_w \sum_d P(w, D) \log_2 \left[\frac{P(w|d)}{P_W(w)} \right] && (671)\\
&= \sum_w \sum_d P(w, D) \log_2 \left[\frac{P(w, d)}{P_W(w) P_D(d)} \right], && (672)
\end{aligned}$$

where we identify the joint distribution of states of the world and data, $P(w, d) = P(w|d) P_D(d)$.

We see that, after all the dust settles, the average 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 $\langle (d \to w) \rangle$.

One consequence of the symmetry or mutuality of information is that we can write the mutual information as a difference of entropies if two different ways,

$$\begin{aligned}
I(d; w) &= S(w) - \sum_d P_D(d) S(w|d) && (673)\\
&= S(d) - \sum_w P_W(w) S(d|w). && (674)
\end{aligned}$$

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

$$\begin{aligned}
I(d; w) &\leq S(w) && (675)\\
I(d; w) &\leq S(d). && (676)
\end{aligned}$$

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. On the other hand, there is a tradition of studying the biological systems as they 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 (676), 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 $P_D(d)$, 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 134: 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.

[I wonder if I should go through the basic calculation of maximum entropy counting here ... since the "things" we count have a cost, this would complete the thought about bounds. At least need a pointer to Appendix A.8.]

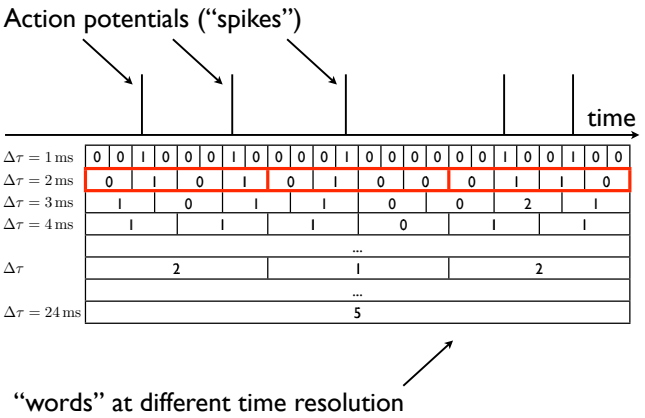

FIG. 131 A schematic of how a train of action potential is converted to discrete "words" at different times resolutions $\Delta\tau$. There is a minimum inter–spike interval, the "refractory period" (here, ~ 2 ms), so that for sufficiently small $\Delta\tau$ the words are binary. Highlighted is the case where $\Delta\tau = 2$ ms and $T = 8$ ms, so this segment of the spike train becomes three successive four bit words, 0101, 0100, and 0110.

To see how the ideas of entropy reduction and information work in a real example, let's consider the response of a neuron to sensory inputs. As we have discussed [starting in Chapter One; give specific pointers], most neurons in the brain generate a sequence of brief (~ 1 ms), identical electrical pulses called action potentials or spikes. Since these events are identical, we can think of them as marking points in time, and then we can build a discrete vocabulary of responses by fixing some limited time resolution $\Delta\tau$, as in Fig 131. More precisely, if $\Delta\tau$ is small, then in each small time window of duration $\Delta\tau$ we will see either one or zero spikes, and so the response is naturally discrete and binary. Then segments of the spike train of duration T can be thought of as $T/\Delta\tau$–letter binary words. Recording from a single neuron as the animal experiences some reasonably complex, dynamic sensory inputs, it is relatively easy to estimate the distribution of these these words, $P(W)$, so long as we don't make the ratio $T/\Delta\tau$ too large. Then we can compute the entropy of this distribution, $S(T, \Delta\tau)$.

Figure 132 shows the results of experiments on the motion sensitive neuron H1 in the fly visual system that we met earlier, in Section [**], when we discussed noise and the precision of visual motion estimation. In these experiments, the fly sees a randomly moving pattern, and H1 responds with a stream of spikes. If we fix $\Delta\tau = 3$ ms and look at $T = 30$ ms segments of the spike train, there are $2^{T/\Delta\tau} \sim 10^3$ possible words, but the distribution is

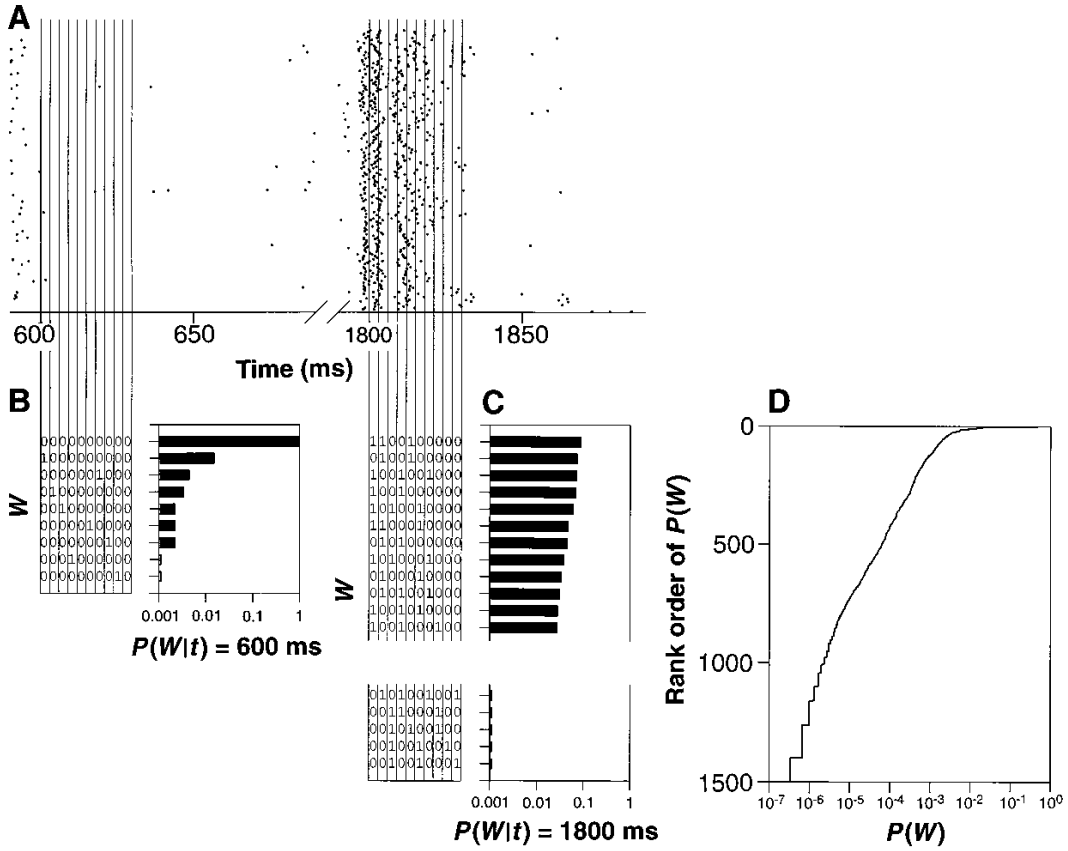

FIG. 132 [Make a new version of this.] A neuron responds to dynamic stimuli with sequences of spikes. In this case, as described in the text, we look at the motion sensitive neuron H1 in the fly's visual system. (A) Each line across time is a single presentation of a movie, and dots mark the arrival times of spikes on each trial. (B) and (C) show the discretization of the spike trains into binary "words" with $\Delta\tau = 3$ ms resolution, and the distribution of words that occur at a particular moment in the movie, $P(w|t)$. (D) The distribution of words averaged over all times, in rank order. From de Ruyter van Steveninck et al (1997).

strongly biased and the entropy is only $S(T, \Delta\tau) \sim 5$ bits. This relatively low entropy means that we can still sample the distributions of words even out to $T \sim 50 - 60$ ms, which is interesting because the fly can actually generate a flight correction in response to visual motion inputs within ~ 30 ms.

The entropy $S(T, \Delta\tau)$ should be an extensive quantity, which means that, for large T, we should have $S(T, \Delta\tau) \propto T$. More strongly, if the correlations in the spike train are sufficiently short ranged, then we expect that at large T we will have

$$\frac{1}{T} S(T, \Delta\tau) = \mathcal{S}(\Delta\tau) + \frac{C(\Delta\tau)}{T} + \cdots , \qquad (677)$$

where $\cdots$ vanish more rapidly than $1/T$. In fact we see this in the real data (Fig 133), which suggests that we really can estimate the entropy rate $\mathcal{S}(\Delta\tau)$.

Connecting to the discussion above, the entropy rate $\mathcal{S}(\Delta\tau)$ sets a limit on the rate at which the spikes can provide information about the sensory input. When we make $\Delta\tau$ smaller, the entropy rate necessarily goes up, because previously indistinguishable responses map to different words at higher time resolution. Concretely, if we make $\Delta\tau$ smaller by a factor of two, then every '1' in the coarse words can become either a '01' or a '10' in the higher resolution words, and so we expect the entropy to increase by roughly one bit for every spike, as in Fig 131.

Problem 135: Entropy and entropy rate in simple models. Going back to Chapter 1, you know how to generate events drawn from a Poisson process with an arbitrary time dependent rate $r(t)$. Here you should take this (semi–)seriously as a model for spike trains, and use the resulting simulations to explore the entropy and entropy rate of neural responses.

(a.) Start with $r = r_0$, a constant. Generate a long sequence of spikes (e.g., $\sim 10^4$). Choose a time resolution $\Delta\tau$ such that $r_0 \Delta\tau \ll 1$, and turn your simulated spike train into a binary sequence; for simplicity ignore the (rare) occurrence of two spikes

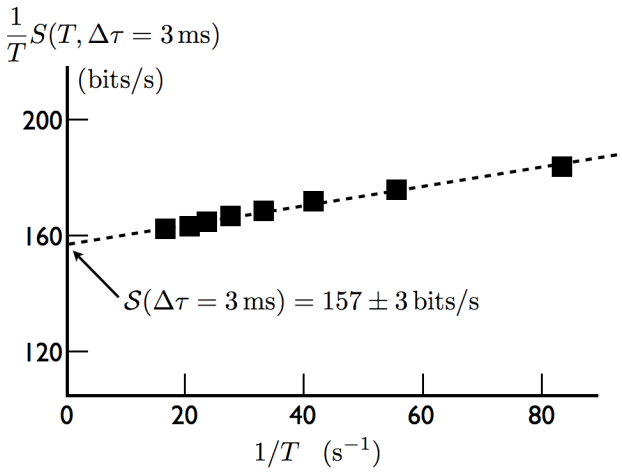

FIG. 133 Entropy is extensive. From the experiments on the neuron H1 in Fig 132, we compute the entropy of words at fixed time resolution $\Delta\tau == 3$ ms and variable length T, stopping when T is so large that we can no longer reliably sample the distribution $P(W)$. The data (error bars are smaller than the symbols) fall on the line predicted in Eq (677), and we can thus extract an estimate of the entropy rate $\mathcal{S}(\Delta\tau)$. Redrawn from Strong et al (1998a).

in one bin. Form "words" with $T/\Delta\tau$ bits, and estimate the distribution of these words from your simulated data. Compute the entropy of this distribution, and explore its dependence on T, r_0, and $\Delta\tau$. Do you see the emergence of an entropy rate, $S \sim \mathcal{S}T$?

(b.) Explain why, for a Poisson process with a constant rate, $S = \mathcal{S}T$ should be exact. From this result, you can calculate $\mathcal{S}$ by thinking about just one bin of size $\Delta\tau$, and you should do this. How does your analytic result compare with the simulation results in (a)?

(c.) Suppose that $x(t)$ is a Gaussian stochastic process with correlation function $\langle x(t)x(t')\rangle = \sigma^2 e^{-|t-t'|/\tau_c}$.[This should be explained somewhere already!] Samples of this process can be generated by simulating the Langevin equation,

$$\tau_c \frac{dx}{dt} = -x + 2\sigma\eta(t), \tag{678}$$

where $\langle \eta(t)\eta(t')\rangle = \delta(t-t')$. Consider a Poisson process with rate $r(t) = r_0 e^{x(t)}$. Generate spike sequences for this process, and follow the procedures in (a) to estimate the entropy in binary words of duration T at resolution $\Delta\tau$, with reasonable choices of parameters. Can you observe the emergence of extensive behavior, $S \sim \mathcal{S}T$? Does this (as seems plausible) require $T \gg \tau_c$? How do your results depend on σ?

A long standing question in thinking about the brain has been whether the precise timing of individual spikes is important, or whether the brain is capable of counting spikes only in relatively coarse time bins, so that the "rate" of spikes over longer periods of time is all that matters. We now have the tools to give a more precise formulation of this question. As we increase our time resolution, the entropy of the spike trains goes up, and hence so does the capacity of the neuron to convey information. The question is whether this capacity is used—does the information about sensory inputs also rise as the time resolution is improved, or is the extra entropy just 'noise'?

If the sensory inputs are called s, then the information that the spike sequences in some window T provide about these inputs can be written, as in Eq (674), as a difference of entropies,

$$I(s;W) = S(W) - \langle S(W|s)\rangle_s, \tag{679}$$

where $\langle\cdots\rangle_s$ denotes an average over the distribution of inputs. We have already discussed the entropy of the neural vocabulary, $S(W)$; the problem is how to estimate $S(W|s)$, the entropy of the words given the sensory input s. To do this we need to sample the distribution $P(W|s)$, that is the distribution of neural responses when the stimulus is fixed. At a minimum, this requires that we repeat the same stimuli many times. So, if the visual stimulus is a long movie, we have to show the movie over and over again. But how do we pick out a particular stimulus s from the continuous stream? One way to do this is to realize that the flow of time in the movie provides an index into the stimuli, and all we need is to be able to compute averages over the distribution of stimuli. If the source of stimuli is ergodic (which we can arrange to be true in the lab!), then an average over stimuli is equivalent to an average over time. So, if we repeat the movie many times, and focus on events at time t relative to the start of the movie, we can sample, in repeats of the movie, the distribution $P(W|t)$, as in Fig 132, and hence estimate $S(W|t)$. Finally, the information is obtained by explicitly replacing the ensemble average with a time average,

$$I(s;W) = S(W) - \langle S(W|t)\rangle_t. \tag{680}$$

Each of the entropy terms on the right should behave as in Eq (677), and so we can extract an estimate of the information rate $R_{\text{info}}(\Delta\tau)$ as a function of time resolution. Results are shown in Fig 134.

We see that, as we vary the time resolution from 800 ms down to 2 ms, the information rate follows the entropy rate, with a nearly constant 50% efficiency. Although we should not generalize too much from one example, this certainly suggests that neurons are making use of a significant fraction of their capacity in actually encoding sensory signals. Also, this is true even at millisecond time resolution. The idea that the entropy of the spike train sets a limit to neural information transmission emerged almost immediately after Shannon's work, but it was never clear whether these limits could be approached by real systems.

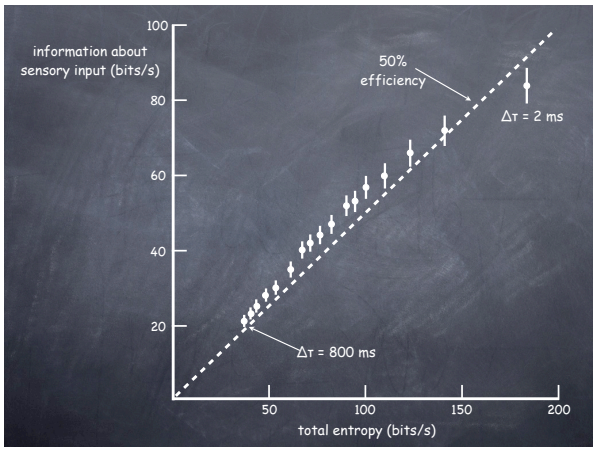

FIG. 134 Entropy and information in a spike train. Experiments on the fly's motion sensitive visual neuron were analyzed as described in the text (following Fig 132) to estimate the total entropy and the information carried about the sensory input. As we vary the time resolution of our analysis from $\Delta\tau = 800\,\text{ms}$ down to $\Delta\tau = 2\,\text{ms}$, we distinguish finer details of the neural response and expand the capacity of the putative neural code; this enhanced capacity is measured by the increasing entropy. Remarkably, across this huge range, capacity is used with almost constant efficiency. From Strong et al (1998a).

Problem 136: Information from single events. This section began by defining the information gained in a single observation. Here, we would like to give the parallel for individual neural responses, but there is a twist because spikes are rare compared with silences. Thus it makes sense to ask how much information we obtain per spike, or per non–silent word W. Imagine that we look in a window of duration $\Delta\tau$ at time t, and we are looking for some event e—this event could be a single action potential, or some combination of multiple spikes with specific intervals between them. On average these events occur with some rate $\bar{r}_e$.

(a.) In the small window $\Delta\tau$, either the event e occurs or it does not; for sufficiently small $\Delta\tau$, the probability of occurrence is $p_e = \bar{r}_e\Delta\tau$. What is the entropy of the binary variable marking the occurrence of the event? Can you simplify your result when $p_e \ll 1$? You'll see that the entropy in this limit is small, but so is the expected number of events. What is the entropy per event?

(b.) If we know the sensory inputs to this neuron, then the probability of an event depends on time, locked to the time dependence of the sensory signal. Let's call the time dependent rate $r_e(t)$. As in (a.), compute the entropy of the binary event/nonevent variable, but now conditional on knowledge of the sensory inputs.

(c.) Combine your results in (a.) and (b.) to give an expression for the mean information that the occurrence or non–occurrence of the event provides about the sensory input. Normalize by the expected number of events, to give bits per event. Is the limit $\Delta\tau \to 0$ well behaved? When the dust settles, you should find that the information per event is

$$I_e = \left\langle \frac{r_e(t)}{\bar{r}_e} \log_2 \left[\frac{r_e(t)}{\bar{r}_e} \right] \right\rangle_t . \tag{681}$$

(d.) As an alternative view of the same question, suppose that we observe a large window of time T. If T is sufficiently large, we can be sure that the event e will occur, but we don't know when.

Problem 137: Information from single spikes in a simple model. In Problem [**] above, you constructed a model spike train using a Poisson process with a time varying rate $r(t) = r_0 e^{x(t)}$, where $x(t)$ is a Gaussian stochastic process. Show that, for this model, the information carried by a single spike about $x(t)$ is linear in the variance of the signal $\langle x^2 \rangle$. This suggests that if the signal variance grows, the information carried by spikes grows with it, without bound. Explain what is wrong with this picture. Suppose instead that the spike rate $r(t)$ depends on x through some saturating function, for example

$$r(t) = \frac{r_0}{1 + \exp[-x(t) + \theta]} . \tag{682}$$

Reduce the formula for I_e in this model to a single integral which you can do numerically. Can you see how the results simplify as $\langle x^2 \rangle$ becomes large? As a hint, notice that this is equivalent to a model in which

$$r(t) = \frac{r_0}{1 + \exp[-\gamma(x(t) + \tilde{\theta})]} , \tag{683}$$

where $\gamma \to \infty$ while $\langle x^2 \rangle$ stays constant. Is there a setting of the threshold θ which maximizes I_e? Is there a cost to achieving this optimum?

One might worry that the high efficiency of coding seen in the fly's H1 neuron arises because the fly has relatively few neurons, and thus is under greater pressure to be efficient. While this may be true, it seems that high coding efficiencies are there to be found even in animals like us and our primate cousins who have very large numbers of neurons. In humans it is possible to record from individual receptor cells in our hands and fingertips, contacting the axons of these cells as they course along the arm to the spinal cord. Data are more limited than in the fly, so one has to be more careful to avoid systematic errors, but the lower bound on the efficiency of coding complex, dynamic variations in the indentation of the skin is above 50%. In the visual cortex of non–human primates, there is a classic series of experiments correlating the perception of motion with the activity of single neurons in area MT. [probably this needs more explanation!] The standard stimuli for these experiments are random dot patterns in which a fraction of the dots move coherently while another fraction are randomly deleted and replaced at new locations; the perception of motion direction becomes less reliable as the degree of coherence decreases. The evidence that single neurons are making a measurable contribution to the perceptual decision is strong, since one can correlate the number of spikes generated by a neuron with the animal's decision about leftward vs. rightward motion even when the coherence is zero, and the animal is just guessing.

The experiments in MT focused on asking the animal to report a decision about motion direction across a two second window of stimulation. When we look at these random patterns, however, we see a certain amount of "jiggling," especially at low coherence. If we present exactly the same pattern of random dots vs. time, we find that the neurons respond with a fair degree of reliability to the temporal details of the movie, certainly down to time scales below 10 ms. In Fig 135 we see what this

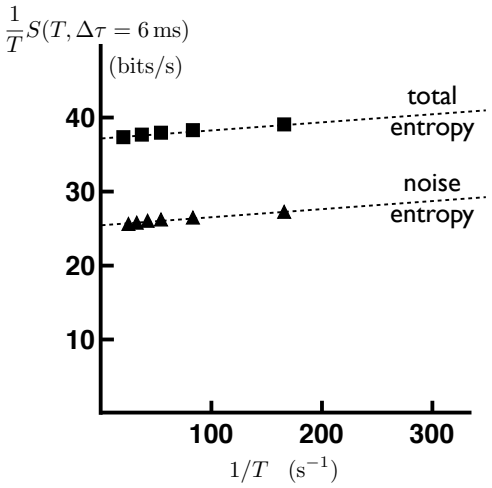

FIG. 135 Entropy and information in spike trains from a motion sensitive neurons in the primate visual cortex (area MT); experiments by Britten et al (1993) and analysis by Strong et al (1998b). [fill in the caption]

means in terms of the information carried by the spike trains about the time–varying details of the visual stimulus, rather than just the overall direction of motion. Here the information, at $\Delta\tau = 6\,\text{ms}$ time resolution, is $25 - 30\%$. Experiments on the same neurons using stimuli that alternated between moving left and right[81] on the $30 - 100\,\text{ms}$ time scale found information rates of $1 - 2.5\,\text{bits/spike}$, quite comparable to the results with H1. In summary, although there are differences in the details of the spike trains from motion sensitive neurons in flies and monkeys, there is little different in the amount of information they carry, or the efficiency with this information is encoded, if we asking about the kinds of complex, dynamic stimuli that are relevant to the real world.

We now want to look at information transmission in the presence of noise, connecting back a bit to what we discussed in Chapters 1 and 2. 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:

$$y = gx + \xi. \tag{684}$$

It seems reasonable in many systems to assume that the noise is Gaussian, either for fundamental physical reasons (as with thermal noise), or because it arises from a superposition of many independent sources, in which case the central limit theorem takes over. 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.

The problem of information transmission with Gaussian signals and noise is sufficiently important that it is worth going through all the algebra quite explicitly; this is also one of those pleasing problems where, as we calculate, terms proliferate and then collapse into a much simpler result. So, onward. The statement that ξ is Gaussian noise means that

$$P(y|x) = \frac{1}{\sqrt{2\pi\langle\xi^2\rangle}} \exp\left[-\frac{1}{2\langle\xi^2\rangle}(y - gx)^2\right]. \tag{685}$$

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

$$P(x) = \frac{1}{\sqrt{2\pi\langle x^2\rangle}} \exp\left[-\frac{1}{2\langle x^2\rangle}x^2\right], \tag{686}$$

and hence y itself is Gaussian,

$$P(y) = \frac{1}{\sqrt{2\pi\langle y^2\rangle}} \exp\left[-\frac{1}{2\langle y^2\rangle}y^2\right] \tag{687}$$

$$\langle y^2\rangle = g^2\langle x^2\rangle + \langle\xi^2\rangle. \tag{688}$$

To compute the information that y provides about x we use Eq. (672):

$$I(y \to x) = \int dy \int dx P(x,y) \log_2\left[\frac{P(x,y)}{P(x)P(y)}\right] \quad \text{bits} \tag{689}$$

$$= \frac{1}{\ln 2}\int dy \int dx P(x,y) \ln\left[\frac{P(y|x)}{P(y)}\right] \tag{690}$$

$$= \frac{1}{\ln 2}\left\langle \ln\left[\frac{\sqrt{2\pi\langle y^2\rangle}}{\sqrt{2\pi\langle\xi^2\rangle}}\right] - \frac{1}{2\langle\xi^2\rangle}(y - gx)^2 + \frac{1}{2\langle y^2\rangle}y^2\right\rangle, \tag{691}$$

where by $\langle\cdots\rangle$ we understand an expectation value over the joint distribution $P(x,y)$. Now in Eq. (691) we can

see that the first term is the expectation value of a 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,

$$I(y \to x) = \frac{1}{\ln 2}\left[\ln\sqrt{\frac{\langle y^2\rangle}{\langle \xi^2\rangle}} - \frac{1}{2} + \frac{1}{2}\right] \tag{692}$$

$$= \frac{1}{2}\log_2\left(\frac{\langle y^2\rangle}{\langle \xi^2\rangle}\right) \tag{693}$$

$$= \frac{1}{2}\log_2\left(1 + \frac{g^2\langle x^2\rangle}{\langle \xi^2\rangle}\right) \quad \text{bits.} \tag{694}$$

Another way of arriving at these results is to remember that the information is a difference of entropies [Eq (674)], but in this case the underlying distributions are all Gaussian. Thus it's useful to know, in general, the entropy of a Gaussian distribution. Suppose that

$$P(z) = \frac{1}{\sqrt{2\pi\langle(\delta z)^2\rangle}}\exp\left[-\frac{(z-\langle z\rangle)^2}{2\langle(\delta z)^2\rangle}\right]. \tag{695}$$

Now our task is to compute

$$S = -\int dz\, P(z)\log_2 P(z) = -\left\langle \log_2 P(z)\right\rangle. \tag{696}$$

But

$$\log_2 P(z) = \frac{1}{\ln 2}\left[\ln\left(\frac{1}{\sqrt{2\pi\langle(\delta z)^2\rangle}}\right) - \frac{(z-\langle z\rangle)^2}{2\langle(\delta z)^2\rangle}\right], \tag{697}$$

and hence

$$S = -\left\langle \log_2 P(z)\right\rangle \tag{698}$$

$$= \frac{1}{\ln 2}\left[\ln\left(\sqrt{2\pi\langle(\delta z)^2\rangle}\right) + \left\langle\frac{(z-\langle z\rangle)^2}{2\langle(\delta z)^2\rangle}\right\rangle\right] \tag{699}$$

$$= \frac{1}{\ln 2}\left[\frac{1}{2}\ln\left(2\pi\langle(\delta z)^2\rangle\right) + \frac{1}{2}\right] \tag{700}$$

$$= \frac{1}{2}\log_2\left[2\pi e\langle(\delta z)^2\rangle\right]. \tag{701}$$

Notice that the entropy is independent of the mean, as we expect, since entropy measures variability or uncertainty.

Problem 138: Using the entropy of Gaussians. Use the general result on the entropy of Gaussian distributions, Eq (701), to rederive Eq (694) for the information transmission through the "Gaussian channel."

We can gain some intuition by rewriting Eq (694). 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:

$$y = g(x + \eta_{\rm eff}), \tag{702}$$

where $\eta_{\rm eff} = \xi/g$. 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, as discussed above.[82] Written in terms of the effective noise level, the information transmission takes a simple form,

$$I(y \to x) = \frac{1}{2}\log_2\left(1 + \frac{\langle x^2\rangle}{\langle \eta_{\rm eff}^2\rangle}\right) \quad \text{bits,} \tag{703}$$

or

$$I(y \to x) = \frac{1}{2}\log_2(1 + SNR), \tag{704}$$

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

The result in Eq. (704) 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 $\delta x_1 \sim \langle \eta_{\rm eff}^2\rangle^{1/2}$, 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 139: 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.

---

[82] As a reminder, 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.

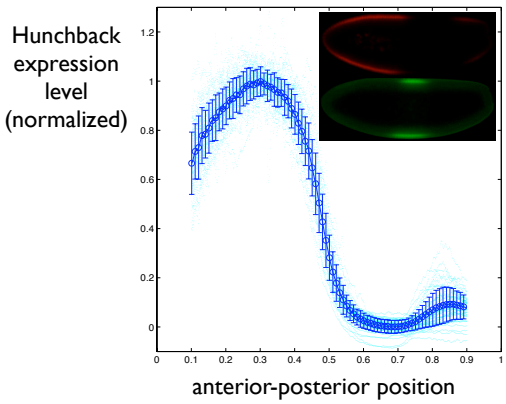

FIG. 136 Spatial profiles of Hunchback expression in the early *Drosophila* embryo. Small dots show experiments from individual embryos; circles with error bars are mean and standard deviation across 51 embryos. In the inset, image in red shows fluorescent antibody staining for Hb, and green shows the corresponding measurement for Krüppel. These images are taken by optical sectioning along the midline of the embryo, and the intensity is measured in a small area, roughly the size of a nucleus, the slides along the "rim" of the embryo where the nuclei are sitting. From Dubuis et al (2011).

To illustrate these ideas, consider the expression of the "gap genes" in the fly embryo, which we have seen in Sections [pointers to specific sections in previous chapters]. We recall that, in response to the primary, maternally supplied morphogens, these genes have varying levels of expression which provide a first step in building the blueprint for the fully developed organism. One of the basic ideas in developmental biology is that these expression levels carry "positional information," i.e. that cells know where they are in the embryo, and hence their fate in the developed organism, as a result of knowing the concentrations of these molecules. It seems natural to ask if we can quantify this positional information, in bits. To do this, as in Fig 136, we can look at many embryos and measure the concentration vs. position in each one. If there is a perfect functional relationship, with no noise, then the transmission of positional information is limited only by the number of samples that we take along the position axis, and hence the information in bits will just be the log of the number of cells. But there is noise, and this sets a limit to the positional information.

The position along the embryo can be measured by $0 \leq x \leq 1$. If we assume that the cells acquiring positional information are distributed uniformly (which is approximately true), then $P(x)$ is uniform, $P(x) = 1$. The expression level of the gene we are looking at will be called g. What we need to know is the distribution of expression levels at one position, $P(g|x)$. Experiments give us samples out of this distribution, but we may or may not have enough samples to characterize the whole distribution. What we can do more easily is to measure the mean $\bar{g}(x)$ and the variance $\sigma_g^2(x)$, and then approximate $P(g|x)$ as being Gaussian. One might worry that this approximation is uncontrolled, but in fact we can say more.

Suppose that all we know is the mean and variance of the distribution $P(g|x)$. The mutual information $I(g;x)$ is the difference between the entropy of the distribution $P(g)$ and the average entropy of the distribution $P(g|x)$,

$$I(g;x) = S[P(g)] - \langle S[P(g|x)]\rangle_x. \tag{705}$$

Thus if we can put an upper bound on the entropy $S[P(g|x)]$, we can put a lower bound on the information. Suppose we search for a distribution $P(g|x)$ that maximizes the entropy, while reproducing the measured mean and variance. As explained in more detail in Appendix [**], we can do this constrained optimization using the standard method of Lagrange multipliers. To maximize $S[P(g|x)]$ we introduce a functional

$$\tilde{S}[P(g|x)] = S[P(g|x)] - \lambda_1 \left[\int dg\, P(g|x) g - \bar{g}(x)\right] - \lambda_2 \left[\int dg\, P(g|x)\, (g - \bar{g}(x))^2 - \sigma_g^2(x)\right]. \tag{706}$$

Now if we maximize $\tilde{S}[P(g|x)]$ with respect to $P(g|x)$, and then extremize with respect to the Lagrange multipliers λ_1 and λ_2, we will find a distribution that maximizes the entropy and reproduces the observed mean and variance. The solution to this problem, as shown in Appendix [**], is the Gaussian distribution. Thus, when we approximate $P(g|x)$ as being Gaussian, we generate a lower bound on the information $I(g;x)$.

In the example of Fig 136, this variance at each position is relatively small, with $\sigma_g(x) \sim 0.1$ in units where the maximum mean expression level is one. Following through the computation of entropies as outlined above, one finds from these data that the expression level of Hunchback protein provides nearly two bits (give exact answer, with error bars) of information about position in the embryo. In the Gaussian approximation this is a lower bound on the information, but in fact the data sets are just large enough to make more direct estimates, and to show that this bound is tight [add a figure to illustrate this]. Classically, the gap genes have been described as

specifying boundaries, dividing the embryo into patches of high (on) and low (off) expression. Evidently a simple on/off picture corresponds at most to one bit of positional information, and so a quantitative analysis teaches us that the focus on "expression boundaries" literally misses half of the story.

Problem 140: Details of positional information. [Develop a problem that asks the students to use some of the real data on the gap genes ...]

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. (684) are then

$$y_i = g_{ij} x_j + \xi_i, \tag{707}$$

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

$$S_{ij} = \langle x_i x_j \rangle \tag{708}$$
$$N_{ij} = \langle \xi_i \xi_j \rangle. \tag{709}$$

The relevant probability distributions are

$$P(\{x_i\}) = \frac{1}{\sqrt{(2\pi)^K \det S}} \exp\left[-\frac{1}{2} x_i \cdot (S^{-1})_{ij} \cdot x_j \right] \tag{710}$$
$$P(\{y_i\}|\{x_i\}) = \frac{1}{\sqrt{(2\pi)^K \det N}} \exp\left[-\frac{1}{2} (y_j - g_{ik} x_k) \cdot (N^{-1})_{ij} \cdot (y_j - g_{jm} x_m) \right], \tag{711}$$

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,

$$P(\{y_i\}) = \int d^K x P(\{y_i\}|\{x_i\}) P(\{x_i\}), \tag{712}$$

and then we write the information as an expectation value,

$$I(\{y_i\} \to \{x_i\}) = \left\langle \log_2 \left[\frac{P(\{y_i\}|\{x_i\})}{P(\{y_i\})} \right] \right\rangle, \tag{713}$$

where $\langle \cdots \rangle$ denotes an average over the joint distribution $P(\{y_i\}, \{x_i\})$. As in Eq. (691), 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

$$I(\{y_i\} \to \{x_i\}) = \frac{1}{2} \mathrm{Tr} \log_2 [\mathbf{1} + N^{-1} \cdot (g \cdot S \cdot g^T)], \tag{714}$$

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

Problem 141: The multi–dimensional Gaussian. Fill in the details leading to Eq (714). [where do I give the problem Tr ln = ln det? connect here]

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." 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 Tr log is equivalent to computing the information transmission along each direction using Eq. (704), 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. (707) we have the continuous linear response of the detector to input signals,

$$y(t) = \int dt' M(t-t')x(t') + \xi(t). \tag{715}$$

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 in Appendix A.2, 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,

$$\langle x(t)x(t')\rangle = C_x(t-t'). \tag{716}$$

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 [pointer] 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. (715), we again refer noise to the input by writing

$$y(t) = \int dt' M(t-t')[x(t') + \eta_{\rm eff}(t')]. \tag{717}$$

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,

$$SNR(\omega) = \frac{S_x(\omega)}{N_{\rm eff}(\omega)}, \tag{718}$$

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

$$I[y(t) \to x(t)] = \frac{1}{2}\sum_\omega \log_2[1 + SNR(\omega)]. \tag{719}$$

Finally, to compute the frequency sum, we recall that [I think this is found also in an Appendix; check!]

$$\sum_n f(\omega_n) \to T\int \frac{d\omega}{2\pi} f(\omega). \tag{720}$$

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

$$I[y(0<t<T) \to x(0<t<T)] \to \frac{T}{2}\int \frac{d\omega}{2\pi}\log_2[1+SNR(\omega)] \text{ bits}. \tag{721}$$

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

$$R_{\rm info} \equiv \lim_{T\to\infty}\frac{1}{T}\cdot I[y(0<t<T)\to x(0<t<T)] \tag{722}$$

$$= \frac{1}{2}\int\frac{d\omega}{2\pi}\log_2[1+SNR(\omega)] \text{ bits/sec}. \tag{723}$$

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 $|\omega|_{\rm max} = \pi/\tau_0$.

Problem 142: 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 learn more by 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 τ, and then jump to something completely new. Given that we integrate for τ, we achieve some signal–to–noise ratio which we'll call $S(\tau)$.

(a.) Explain why, in this simple model, if the noise is Gaussian then the rate at which we gain information is at most

$$R_{\rm info}(\tau) = \frac{1}{\tau}\log_2[1+S(\tau)]. \tag{724}$$

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 receptor responds to its inputs according to a simple differential equation

$$\tau_0\frac{dy}{dt} = -y + x. \tag{725}$$

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

$$\langle \eta_{\rm eff}(t)\eta_{\rm eff}(t')\rangle = \sigma_0^2 e^{-|t-t'|/\tau_c}. \quad (726)$$

Show that if we average the noise over a window of duration τ, then the variance

$$\sigma^2(\tau) \equiv \left\langle \left[\frac{1}{\tau}\int_0^\tau dt\, \eta_{\rm eff}(t)\right]^2 \right\rangle \approx \sigma_0^2 \quad (\tau \ll \tau_0) \quad (727)$$

$$\approx \frac{2\sigma_0^2\tau_c}{\tau} \quad (\tau \gg \tau_0). \quad (728)$$

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 τ_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 $R_{\rm info}(\tau)$; show that with proper choice of units, you don't need to know the value of A. What value of τ 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 $dR_{\rm info}/d\tau = 0$. 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 τ is varied? Use this to bound $S(\tau)$ at the point of maximum $R_{\rm info}$. 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]?

In the same way that we used the Gaussian approximation to put bounds on the positional information carried by the gap genes, we can put bounds on the information carried by sensory neurons. As discussed in Section [**], we can reconstruct continuous sensory input signals from the discrete sequences of action potentials, sometimes quite accurately. Concretely, the sensory stimulus $s(t)$ could be light intensity as a function of time in a small region of the visual field, sound pressure as a function of time at the ear canal, the amplitude of mechanical vibrations in sensors such as the cricket cercus and frog sacculus, ... We can estimate the signal from the spike times $\{t_{\rm i}\}$ in a single neuron as

$$s_{\rm est}(t) = \sum_{\rm i} f(t - t_{\rm i}), \quad (729)$$

where the filter $f(\tau)$ is chosen to minimize $\chi^2 = \langle |s_{\rm est}(t) - s(t)|^2\rangle$. Then the quality of the reconstructions can be evaluated by measuring the power spectrum of errors in the reconstruction, and referring these errors to the input, frequency component by frequency component,

$$\tilde{s}_{\rm est}(\omega) = g(\omega)\left[\tilde{s}(\omega) + \tilde{\eta}_{\rm eff}(\omega)\right]. \quad (730)$$

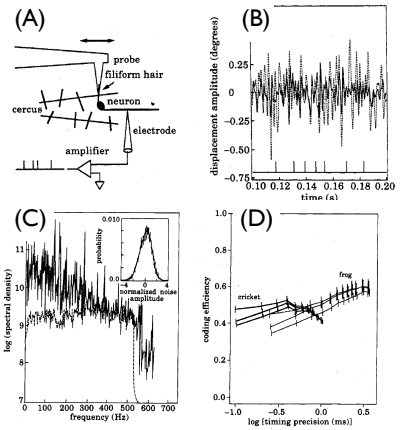

FIG. 137 Coding efficiency in cricket and frog vibration sensors. (A) A schematic of experiments on the cricket cercal sensors, with direct stimulation of the sensory hairs and recording from the primary sensory neurons. (B) Stimulus (dashed) and reconstruction (Solid line) in experiments on the cercal neurons. (C) Power spectral density of the signal, and the noise $\eta_{\rm eff}$ in the reconstructions, from Eq(730). (D) Coding efficiency for example neurons in the cricker cercus and the frog sacculus, using successively higher order approximations to the spike train entropy. Variable timing precision is implemented by providing the reconstruction algorithm in Eq (729) with spike times $t_{\rm i}$ at limited resolution. From Rieke et al (1993).

Although the errors in the reconstruction might not be exactly Gaussian, the maximum entropy argument above tells us that we can put a lower bound on the information which the spike train provides about the stimulus $s(t)$ by measuring the power spectrum of the effective noise $\eta_{\rm eff}$. An example is shown in Fig 137, from experiments on the mechanical sensors in the cricket and frog. Importantly, we can also put upper bounds on the entropy of the spike train, first by assuming that spikes occur independently, then by assuming that the intervals between spikes are independent, then allowing for correlations between successive intervals. With a lower bound on the information and an upper bound on the entropy, we have a lower bound on their ratio, the coding efficiency. In these systems, as with the case of H1 in Fig 134, we see that efficiencies reach $\sim 50\%$ with timing precision in the millisecond range.

By now both the "direct" and the "reconstruction" methods have been used to measure information rates and coding efficiencies in a wide range of neurons responding to sensory stimuli, from the first steps of sensory coding in invertebrates, such as the cricket cercal system in Fig 137, to cells deep in primate visual cortex. The result that single neurons use 30–50% of their spike train entropy to encode sensory information, even down to millisecond resolution, has been confirmed in many systems [maybe reminder that references are at the end of the section?]. An important thread running through

this work is that information rates and coding efficiencies are higher, and the high coding efficiency extends to higher time resolution, when sensory inputs are more like those which occur in nature—complex, dynamic, and with enormous dynamic range; an example from the frog auditory system is shown in Fig 138 [do we need more examples here?]. These results suggest not only that the brain is capable of efficient coding, but also that this efficiency is achieved by matching neural coding strategies to the structure of natural sensory inputs. We will return to this idea in Section IV.C.

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, $N_{\rm eff}(\omega)$. 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, 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.

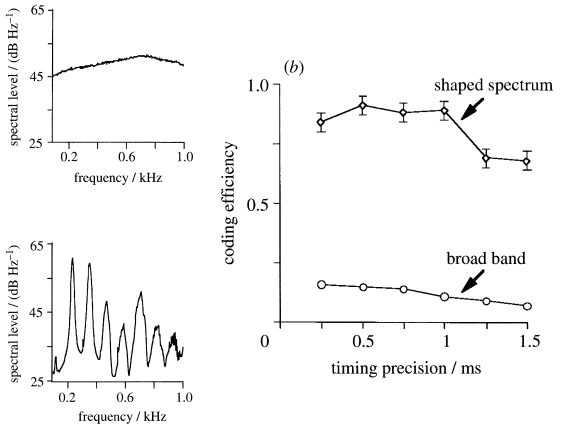

FIG. 138 Coding efficiency in frog auditory neurons. At left, the power spectrum of a broadband, artificial stimulus (top) and a stimulus shaped to have the same spectrum as bullfrog calls (bottom). These stimuli were played to the bullfrog while recording from individual auditory neurons emerging from the amphibian papilla. Reconstructing the sound pressure as a function of time allows us to bound the information transmission rate, as explained in the text, and from this we estimate the coding efficiency—the ratio of the information rate to the entropy rate. In this example, at right, we see clearly that the coding efficiency is substantially higher for the more naturalistic stimuli, approaching 90%. From Rieke et al (1995).

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

$$\tilde{R} = R_{\rm info} - \lambda\langle x^2\rangle \tag{731}$$

$$= \frac{1}{2}\int \frac{d\omega}{2\pi}\log_2\left[1+\frac{S_x(\omega)}{N_{\rm eff}(\omega)}\right] - \lambda\int\frac{d\omega}{2\pi}S_x(\omega). \tag{732}$$

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

$$\frac{\delta\tilde{R}}{\delta S_x(\omega)} = \frac{1}{2\ln 2}\cdot\frac{1}{1+S_x(\omega)/N_{\rm eff}(\omega)}\cdot\frac{1}{N_{\rm eff}(\omega)} - \lambda, \tag{733}$$

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

$$S_x(\omega) + N_{\rm eff}(\omega) = \frac{1}{2\lambda\ln 2}. \tag{734}$$

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. (734) 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. (734) 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 143: Whitening. Consider a system that responds linearly to a signal $s(t)$, with added noise $\eta(t)$:

$$x(t) = \int d\tau\, F(\tau)s(t-\tau) + \eta(t). \tag{735}$$

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

$$\langle\eta(t)\eta(t')\rangle = \mathcal{N}_0\delta(t-t'). \tag{736}$$

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

$$\langle s(t)s(t')\rangle = \int \frac{d\omega}{2\pi} S(\omega)\exp[-i\omega(t-t')]. \quad (737)$$

(a.) Write an expression for the rate $R_{\rm info}$ 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 (735) but with $\eta = 0$. Find an expression for the variance of this "ouput signal."

(c.) Consider the problem of maximizing $R_{\rm info}$ 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 becomes 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

$$S(\omega) = \frac{2\langle s^2\rangle\tau_{\rm c}}{1+(\omega\tau_{\rm c})^2}, \quad (738)$$

where $\tau_{\rm c}$ is the correlation time. Find the optimal filter in this case, and use this to evaluate the maximum value of $R_{\rm info}$ as a function of the output signal variance. You should check that your results for $R_{\rm info}$, 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. We have seen these data before, in thinking about how the precision of photon counting changes as the background light intensity increases. Recall from Section I.A that, over a reasonable dynamic range of intensity variations, 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:

$$V(t) = V_{\rm DC} + \int dt' T(t-t')C(t') + \delta V(t). \quad (739)$$

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,

$$V(t) = V_{\rm DC} + \int dt' T(t-t')[C(t') + \delta C_{\rm eff}(t)]. \quad (740)$$

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,

$$\langle \delta C_{\rm eff}(t)\delta C_{\rm eff}(t')\rangle = \int \frac{d\omega}{2\pi} N_C^{\rm eff}(\omega)\exp[-i\omega(t-t')]. \quad (741)$$

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 noise into estimates of the information capacity of these cells. This was done both for the photoreceptor cells and for the large monopolar cells (LMCs) 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 139.

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 the 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 a six times higher rate. Finally, in the large monopolar cells in-

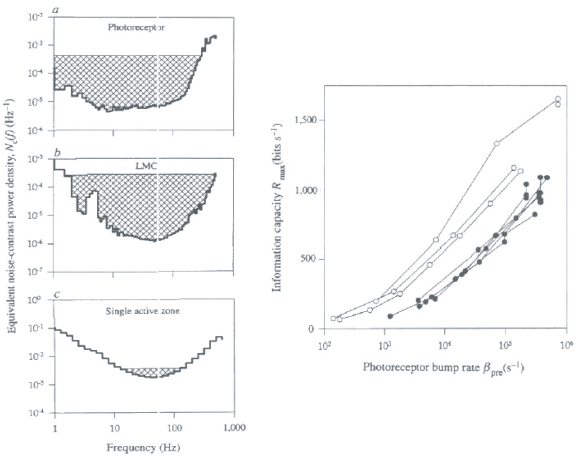

FIG. 139 At left, the effective contrast noise levels in a single photoreceptor cell, a single LMC (the second order cell) and the inferred noise level for a single active zone of the synapse from photoreceptor to LMC. The hatching shows the signal spectra required to whiten the total output over the 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 (1996).

formation 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."

[This discussion needs to be fleshed out. It's also the second independent use of max ent in this section, which makes me worry that leaving max ent to an Appendix may be a mistake, although it also comes up earlier .. this is a pretty big organizational isse. Also was thinking of being explicit about max ent for counting, above, which would make things easier here! 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 (see, for example, Appendix A.8). 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.

[This section needs a summary and conclusion!]

To a remarkable extent, Shannon's original work provides a complete and accessible guide to the foundations of the subject (Shannon 1948). Seldom has something genuinely new emerged so fully in one (admittedly long, two part) paper. For a modern textbook account, the standard is set by Cover & Thomas (1991). An fascinating if idiosyncratic treatment of Shannon's ideas is given by Brillouin (1962). A recent textbook that emphasizes connections between information theory and statistical physics is Mézard & Montanari (2009). The brief discussion of four letter words is based on Stephens & Bialek (2010).

Brillouin 1962: *Science and Information Theory* L Brillouin (Academic, New York, 1962).

Cover & Thomas 1991: *Elements of Information Theory* TM Cover & JA Thomas (Wiley, New York, 1991); there is also a second edition (2006).

Mézard & Montanari 2009: *Information, Physics and Computation.* M Mézard & A Montanari (Oxford University Press, Oxford, 2009).

Shannon 1948: A mathematical theory of communication, CE Shannon, *Bell Sys. Tech. J.* **27,** 379–423 & 623–656 (1948). Reprinted in CE Shannon & W Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Urbana, 1949).

Stephens & Bialek 2010: Statistical mechanics of letters in words. GJ Stephens & W Bialek, *Phys Rev E* **81,** 066119 (2010); arXiv:0801.0253 [q–bio.NC] (2008).

The exploration of neural coding using ideas from information theory rests on a large literature, starting with Adrian's first experiments recording the spikes from individual sensory neurons in the 1920s. For a guide to the field up to the mid 1990s, see Rieke et al (1997). The idea of using ergodicity to make "direct" estimates of the entropy and information in spike trains as they encode dynamic signals is presented by Strong et al (1998a) and de Ruyter van Steveninck et al (1997). For the original discussion of the limits to information transmission by spike trains, see MacKay & McCulloch (1952). For the analysis of information carried by single events, see DeWeese & Meister (1999) and Brenner et al (2000).

Brenner et al 2000: Synergy in a neural code. N Brenner, SP Strong, R Koberle, W Bialek & RR de Ruyter van Steveninck, *Neural Comp* **12,** 1531–1552 (2000); arXiv:physics/9902067 (1999).

DeWeese & Meister 1999: How to measure the information gained from one symbol. MR DeWeese & M Meister, *Network* **10,** 325–340 (1999).

MacKay & McCulloch 1952: The limiting information capacity of a neuronal link. D MacKay & WS McCulloch, *Bull Math Biophys* **14,** 127–135 (1952).

de Ruyter van Steveninck et al 1997: Reproducibility and variability in neural spike trains. RR de Ruyter van Steveninck, GD Lewen, SP Strong, R Koberle & W Bialek, *Science* **275,** 1805–1808 (1997).

Strong et al 1998a: Entropy and information in neural spike trains. SP Strong, R Koberle, RR de Ruyter van Steveninck & W Bialek, *Phys Rev Lett* **80,** 197–200 (1998); arXiv:cond–mat/9603127 (1996).

The experiments connecting motion perception to the activity of individual neurons in visual cortex were reported in a series of beautiful papers by Newsome, Movshon, and their collaborators (Newsome et al 1989) [give a proper guide!]. Of particular relevance here is Britten et al (1993). The fact that these experiments generated demonstrated reproducible responses to stimuli that repeated their temporal details was emphasized by Bair & Koch (1996). The

analysis in Fig 135 is from Strong et al (1998b), unpacking a footnote in Strong et al (1998a) above. Experiments designed to look more specifically at these problems of information transmission for dynamic signals in MT were done by Buračas et al (1998).

Bair & Koch 1996: Temporal precision of spike trains in extrastriate cortex of the behaving macaque monkey, W Bair & C Koch, *Neural comp* **8,** 1185–1192 (1996).

Britten et al 1993: Response of neurons in macaque MT to stochastic motion signals. KH Britten, MN Shadlen, WT Newsome & JA Movshon, *Vis Neuroci* **10,** 1157–1169 (1993).

Buračas et al 1998: Efficient discrimination of temporal patterns by motion–sensitive neurons in primate visual cortex. GT Buračas, AM Zador, MR DeWeese & TD Albright, *Neuron* **20,** 959–969 (1998).

Newsome et al 1989: Neuronal correlates of a perceptual decision. WT Nesome, KH Britten & JA Movshon, *Nature* **341,** 52–54 (1989).

Strong et al 1998b: On the application of information theory to neural spike trains. SP Strong, RR de Ruyter van Steveninck, W Bialek & R Koberle, in *Pacific Symposium on Biocomputing '98*, RB Altman, AK Dunker, L Hunter & TE Klein, eds, pp 621–632 (World Scientific, Singapore, 1998).

The classic discussion of information transmission in the presence of Gaussian noise is again by Shannon (1949), and again the standard modern textbook account is in Cover and Thomas (1991), cited in Section IV.A. The discussion of positional information is based on Dubuis et al (2011). The idea of decoding the spike train to recover the underlying signal ("stimulus reconstruction") was introduced in Section [pointer], and is reviewed by Rieke et al (1997), and the first use of this approach to compare information and entropy was by Rieke et al (1993). This led to experiments showing that coding efficiencies are higher for more naturalistic stimuli (Rieke et al 1995), and this problem was eventually revisited in the context of much more natural stimuli using the "direct" methods of information estimation coupled with more sophisticated strategies for dealing with finite sampling Nemenman et al (2008), as explained in Appendix A.9. The measurements on information capacity in the fly retina are by de Ruyter van Steveninck and Laughlin (1996),. The comparison of this information rate with the limits set by counting vesicles is discussed by Rieke et al (1997), above [probably need to unpack the original refs about vesicle counting rates].

Dubuis et al 2011: Positional information, in bits. JO Dubuis, G Tkačik, W Bialek, EF Wieschaus & T Gregor, in preparation (2011).

Nemenman et al 2008: Neural coding of a natural stimulus ensemble: Information at sub–millisecond resolution. I Nemenman, GD Lewen, W Bialek & RR de Ruyter van Steveninck, *PLoS Comp Bio* **4,** e1000025 (2008); arXiv:q–bio.NC/0612050 (2006).

Rieke et al 1993: Coding efficiency and information rates in sensory neurons. F Rieke, D Warland & W Bialek, *Europhys. Lett* **22,** 151–156 (1993).

Rieke et al 1995: Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory neurons. F Rieke, DA Bodnar & W Bialek, *Proc R. Soc Lond. Ser. B* **262,** 259–265 (1995).

Rieke et al 1997: *Spikes: Exploring the Neural Code* F Rieke, D Warland, RR de Ruyter van Steveninck & W Bialek (MIT Press, Cambridge, 1997).

de Ruyter van Steveninck & Laughlin 1996: The rate of information transfer at graded–potential synapses. RR de Ruyter van Steveninck & SB Laughlin, *Nature* **379,** 642–645 (1996).

Shannon 1949: Communication in the presence of noise. CE Shannon, *Proc IRE* **37,** 10–21 (1949).

I need to add a guide to all the experiments on information rates and coding efficiencies, using both direct and reconstruction methods. Even this list is incomplete.

Attias & Schreiner 1998: Coding of naturalistic stimuli by auditory midbrain neurons. H Attias & CE Schreiner, in *Advances in Neural Information Processing Systems 10*, MI Jordan, MJ Kearns & SA Solla, ppp 103–109 (MIT Press, Cambridge, 1998).

Bair & Koch 1996: Temporal precision of spike trains in extrastriate cortex of the behaving macaque monkey. W Bair & C Koch, *Neural Comp* **8,** 44–66 (1996).

Buračas et al 1998: Efficient discrimination of temporal patterns by motion–sensitive neurons in primate visual cortex. GT Buračas, AM Zador, MR DeWeese & TD Albright, *Neuron* **20,** 959–969 (1998).

Escabi et al 2003: Naturalistic auditory contrast improves spectrotemporal coding in the cat inferior colliculus. MA. Escab, LM Miller, HL Read & CE Schreiner, *J Neurosci* **23,** 11489–11504 (2003).

Kara et al 2000: Low response variability in simultaneously recorded retinal, thalamic, and cortical neurons. P Kara, P Reinagel & RC Reid, *Neuron* **27,** 635–646 (2000).

Koch et al 2006: How much he eye tells the brain. K Koch, J McLean, R Segev, MA Freed, MJ Berry II, V Balasubramanian & P Sterling, *Curr Biol* **16,** 1428–1434 (2006).

Liu et al 2001: Variability and information in a neural code of the cat lateral geniculate nucleus. RC Liu, S Tzonev, S Rebrik & KD Miller, *J Neurophysiol* **86,** 2789–2806 (2001).

Reinagel & Reid 2000: Temporal coding of visual information in the thalamus. P Reinagel & RC Reid, *J Neurosci* **20.** 5392–5400 (2000).

Rokem et al 2006: Spike–timing precision underlies the coding efficiency of auditory receptor neurons. A Rokem, S Watzl, T Gollisch, M Stemmler, AVM Herz & I Samengo, *J Neurophysiol* **95,** 2541–2552 (2006).

Simmons & de Ruyter van Steveninck 2010: Sparse but specific temporal coding by spikes in an insect sensory–motor ocellar pathway. PJ Simmons & RR de Ruyter van Steveninck, *J Exp Biol* **213,** 2629–2639 (2010).

Yu et al 2005: Preference of sensory neural coding for $1/f$ signals. Y Yu, R Romero & TS Lee, *Phys Rev Lett* **94,** 108103 (2005).

I haven't said anything about error correcting codes. I don't see, in the short run, how to connect these elegant ideas to real biological phenomena. On the other hand, they are so interesting ... at the very least I will need to give references, and some commentary about why we should be trying to think about this.

B. Does biology care about bits?

The question for this section has been with us almost since Shannon's original work. One the one hand, the few examples we have seen in the last section certainly suggest that organisms are squeezing more bits out of

their hardware than we might naively have expected, perhaps even coming close to physical limits on information transmission. On the other hand, the usual view of information theory is as a theory for communication, with its most sophisticated developments in the context of error correcting codes, which seem of little relevance to the natural (as opposed to the engineered) world. Here we'll review old ideas about the connection of information to gambling, and see how closely related ideas have reappeared in thinking about the life strategies of bacterial populations. Then we'll step back and try to look more generally at the connections among information, biological function and evolutionary fitness, and argue that evolution really can select for biological mechanisms that are efficient in an information theoretic sense.

To start, let us consider a simple game; this may seem like a strange topic for a physics course, but please bear with me! I will flip a coin, and you bet on whether it will come up heads or tails. If you get it right, I double your money. If you're wrong, you lose what you bet. If this is a fair coin, so that heads and tails each come up half the time, there really isn't anything to analyze, what happens is "just chance." But if you know, for example, that this is a biased coin, and that the probability of heads really is 60%, you might be tempted to put all of your money on heads. On average, if you bet one dollar you will receive $2 \times (0.6) = 1.2$ dollars in return, which sounds good. Indeed, if we play only once then this is what you should do, since it will maximize your expected return.

But what happens if we are going to play repeatedly, which you might think is a better metaphor for life? Now if you put all your money on heads, there is a 40% chance that, in one flip, you'll lose it all. Suppose that instead you put a fraction f of your money on heads and a fraction $1-f$ on tails. If we introduce a binary variable $n = 1$ for heads and $n = 0$ for tails, then on the i^{th} flip your winnings will change by a factor

$$G_{\mathrm{i}} = 2 \times \left[f n_{\mathrm{i}} + (1-f)(1-n_{\mathrm{i}}) \right], \tag{742}$$

where n_{i} marks what happens on the i^{th} flip. After N successive flips you will have a gain

$$G_{\mathrm{total}}(N) = 2^N \prod_{\mathrm{i}=1}^{N} \left[f n_{\mathrm{i}} + (1-f)(1-n_{\mathrm{i}}) \right], \tag{743}$$

where we are assuming that you consistently put a fraction f of your accumulated winnings down as a bet on heads, and the remainder on tails.

To keep going, we want to write the product in Eq (743) as the exponential of a sum. It's useful to notice that, becuase n_{i} is either 0 or 1, we have

$$f n_{\mathrm{i}} + (1-f)(1-n_{\mathrm{i}}) = \exp\left[n_{\mathrm{i}} \ln(f) + (1-n_{\mathrm{i}}) \ln(1-f) \right]. \tag{744}$$

This means that we can write the total gain

$$\begin{aligned} G_{\mathrm{total}}(N) &= 2^N \prod_{\mathrm{i}=1}^{N} \left[f n_{\mathrm{i}} + (1-f)(1-n_{\mathrm{i}}) \right] \\ &= 2^N \prod_{\mathrm{i}=1}^{N} \exp\left[n_{\mathrm{i}} \ln(f) + (1-n_{\mathrm{i}}) \ln(1-f) \right] \end{aligned} \tag{745}$$

$$= \exp\left[N \Lambda(f; \{n_{\mathrm{i}}\}) \right], \tag{746}$$

where

$$\Lambda(f; \{n_{\mathrm{i}}\}) = \ln 2 + \frac{1}{N} \sum_{\mathrm{i}=1}^{N} \left[n_{\mathrm{i}} \ln(f) + (1-n_{\mathrm{i}}) \ln(1-f) \right] \tag{747}$$

Written this way, $\Lambda(f; \{n_{\mathrm{i}}\})$ define a rate of exponential growth for your winnings. But $\Lambda(f; \{n_{\mathrm{i}}\})$ depends not only on your betting strategy, summarized by the fraction f that you put on heads, but also on the sequence of heads and tails that come up in the game, denoted by $\{n_{\mathrm{i}}\}$. The key point is that, if we play *many* times, so we can think about the limit $N \to \infty$, this dependence on the details of the flips goes away.

We recall that, for any well behaved random variable, the average over N observations must approach the mean computed from the probability distribution as N becomes large. In the present case, if n_{i} is a binary variable that takes the value $n_{\mathrm{i}} = 1$ with probability p and $n_{\mathrm{i}} = 0$ with probability $1-p$, then as N becomes large we should have

$$\frac{1}{N} \sum_{\mathrm{i}=1}^{N} n_{\mathrm{i}} \to p, \tag{748}$$

and similarly

$$\frac{1}{N} \sum_{\mathrm{i}=1}^{N} (1-n_{\mathrm{i}}) \to 1-p. \tag{749}$$

We can use this to evaluate the long term growth of your winnings, simplifying the results of Eq (747):

$$\frac{1}{N}\ln G_{\rm total}(N) \equiv \Lambda(f) = \ln 2 + \frac{1}{N}\sum_{\rm i=1}^{N}[n_{\rm i}\ln(f) + (1-n_{\rm i})\ln(1-f)] \quad (750)$$

$$= \ln 2 + \left(\frac{1}{N}\sum_{\rm i=1}^{N} n_{\rm i}\right)\ln(f) + \left(\frac{1}{N}\sum_{\rm i=1}^{N}(1-n_{\rm i})\right)\ln(1-f)$$

$$\to \ln 2 + p\ln(f) + (1-p)\ln(1-f), \quad (751)$$

where again p is the probability of heads. To maximize the growth rate $\Lambda(f)$, as usual we differentiate and set the result to zero:

$$\Lambda(f) = \ln 2 + p\ln(f) + (1-p)\ln(1-f)$$

$$\frac{d\Lambda(f)}{df} = p\frac{1}{f} + (1-p)(-1)\frac{1}{1-f}; \quad (752)$$

$$\left.\frac{d\Lambda(f)}{df}\right|_{f=f_{\rm opt}} = 0$$

$$\Rightarrow 0 = p\frac{1}{f_{\rm opt}} + (1-p)(-1)\frac{1}{1-f_{\rm opt}} \quad (753)$$

$$\frac{1-p}{1-f_{\rm opt}} = \frac{p}{f_{\rm opt}}, \quad (754)$$

or more simply $f_{\rm opt} = p$. This is an interesting result: you maximize the rate at which your winnings will grow by "matching" the fraction of your resources that you bet on heads to the probability that the coin will come up heads, and similarly for tails.

Problem 144: Check that $f_{\rm opt} = p$ is a maximum, and not a minimum, of $\Lambda(f)$.

Problem 145: If we bet only once, then in this simple game the maximum mean payoff is obtained by betting on the most likely outcome. On the other hand, as we play many times—more precisely, in the limit that we play infinitely many times—what we have seen is that a sort of matching strategy, or "proportional gambling" maximizes the growth rate. Explore the crossover between these limits. You might start with some simple simulations, and then see if you can make analytic progress, perhaps saying something about the leading $1/N$ corrections at large N. I am leaving this deliberately vague and open ended, hoping that you will play around.

Something even more interesting happens when we evaluate the optimal growth rate, that is $\Lambda_{\rm opt} = \Lambda(f_{\rm opt})$:

$$\Lambda_{\rm opt} = \Lambda(f = p) \quad (755)$$

$$= \ln 2 + p\ln(p) + (1-p)\ln(1-p) \quad (756)$$

$$= \ln 2 - [-p\ln(p) - (1-p)\ln(1-p)]\,. \quad (757)$$

These terms should be starting to look familiar. The term $\ln 2$ is the entropy for a binary variable (heads/tails) if you don't know anything about what to expect, and hence the two alternatives are equally likely. In contrast, the term in brackets,

$$-p\ln(p) - (1-p)\ln(1-p),$$

is the entropy of a binary variable if you know that the two alternatives come up with probabilities p and $1-p$. Thus the optimal growth rate is the difference in entropy between what might happen with an arbitrary coin and what you know will happen with this coin. In other words, *the maximum rate at which your winnings can grow in a simple gambling game is equal to the information that you have about the outcome of a single coin flip.*

This connection between information theory and gambling was discovered in the 1950s by Kelly, who was searching for some interpretation of Shannon's work that didn't refer to the process of communication. Obviously what we have worked out here is a very simple and special case, and we need to do much more in order to claim that the connection is general. But before launching into this let me emphasize something about Kelly's result. At some intuitive level, we can all agree that if we know more about the outcome of the coin flip (or the horse race, or the stock market, or ...) then we should be able to make more money. In a very general context, Shannon proved that "know more" should be quantified by various entropy–like quantities, but it's not obvious that the knowledge measured by Shannon's bits is actually the useful knowledge when it comes time to make a bet. Even if bits are the right measure, the connection between information and the growth of winnings could have been much more vague; you could imagine, for example, that the growth rate is bounded by some function of the information, and that this bound might or might not be realizable with feasible strategies. In contrast to these pessimistic alternatives, Kelly showed that the maximum growth rate *is* the information, and his proof is constructive so we actually know how to achieve this maximum. This really is quite astonishing.

Let's try to generalize what what we have done. Suppose that on each trial i, there are many possible outcomes, $\mu = 1, 2, \cdots, K$; we'll write $n_{\rm i}^{(\mu)} = 1$ if on the $\rm i^{th}$ trial the outcome is μ, and $n_{\rm i}^{(\mu)} = 0$ otherwise. Further, let's say that you bet a fraction of your assets f_μ on each of the possible outcomes μ, and if μ actually happens then each dollar bet on this outcome becomes g_μ dollars;

all money bet on things that don't happen is lost. If you need an example of this sort of game, think of a horse race in which you get something back only if you pick the winner. We'll assume that the different outcomes occur with probability p_μ, but we won't assume anything about the relationship between these odds and the payoffs g_μ.

Having defined all the factors, the analog of Eq (743), is

$$G_{\text{total}}(N) = \prod_{i=1}^{N} \left[\sum_{\mu=1}^{K} f_\mu g_\mu n_i^{(\mu)} \right]. \tag{758}$$

Now we can follow the same steps as before:

$$\ln G_{\text{total}}(N) = \sum_{i=1}^{N} \ln \left[\sum_{\mu=1}^{K} f_\mu g_\mu n_i^{(\mu)} \right] \tag{759}$$

$$= \sum_{i=1}^{N} \sum_{\mu=1}^{K} n_i^{(\mu)} \ln(f_\mu g_\mu) \tag{760}$$

$$\frac{1}{N} \ln G_{\text{total}}(N) = \sum_{\mu=1}^{K} \left[\frac{1}{N} \sum_{i=1}^{N} n_i^{(\mu)} \right] \ln(f_\mu g_\mu) \tag{761}$$

$$\to \Lambda(\{f_\mu\}) = \sum_{\mu=1}^{K} p_\mu \ln(f_\mu g_\mu). \tag{762}$$

We want to maximize the growth rate Λ, subject to the normalization condition that the fractions of our assets placed on all the options add up ($\sum_\mu f_\mu = 1$), so we introduce a Lagrange multiplier α and find the maximum of the function

$$\tilde{\Lambda}(\{f_\mu\}) = \sum_{\mu=1}^{K} p_\mu \ln(f_\mu g_\mu) - \alpha \left[\sum_{\mu=1}^{K} f_\mu - 1 \right]. \tag{763}$$

The equations for the maximum are, as usual,

$$\left. \frac{\partial \tilde{\Lambda}(\{f_\mu\})}{\partial f_\mu} \right|_{\{f_\mu\}=\{f_\mu^{\text{opt}}\}} = 0 \tag{764}$$

$$\Rightarrow 0 = \frac{p_\mu}{f_\mu^{\text{opt}}} - \alpha, \tag{765}$$

$$f_\mu^{\text{opt}} = \frac{p_\mu}{\alpha}; \tag{766}$$

since $\sum_\mu f_\mu = \sum_\mu p_\mu = 1$, we must have $\alpha = 1$, so that

$$f_\mu^{\text{opt}} = p_\mu. \tag{767}$$

Substituting, we find the maximum growth rate

$$\Lambda_{\text{opt}} = \sum_{\mu=1}^{K} p_\mu \ln(p_\mu g_\mu). \tag{768}$$

The first interesting thing is that we recover from the simpler heads/tails problem the idea of proportional gambling [Eq (767)]: you maximize the rate at which your winnings will grow by "matching" the fraction of your resources that you bet on each horse in the race to the probability that this horse will win. Strangely, this is independent of the rewards or gains as expressed in the parameters $\{g_\mu\}$.

[At some point should make a connection between proportional gambling and "matching" behavior .. is this understood?]

The second point is that we can see what it means for the odds to be truly fair. If our opponent in this game (the track operator) sets the returns in inverse proportion to the probability that each horse wins, $g_\mu = 1/p_\mu$, then the maximum growth rate of our winnings, Λ_{opt}, is exactly zero.

This notion of fairness leads us to an information theoretic interpretation of Λ_{opt}. Notice that we have done our calculation on the assumption that we have perfect knowledge of the distribution $\{p_\mu\}$. Perhaps the track operators have less knowledge, and so they set the odds *as if* the distribution were something else, which we can call $\{q_\mu\}$. More generally, we can define

$$q_\mu = \frac{1}{Z} \frac{1}{g_\mu}, \tag{769}$$

with Z chosen so that $\sum_\mu q_\mu = 1$. If $Z = 1$, then the payoffs $\{g_\mu\}$ are fair in the distribution $\{q_\mu\}$, while if $Z < 1$ the track operators are keeping something for themselves (as they are wont to do). Then we can see that

$$\Lambda_{\text{opt}} = -\ln Z + \sum_{\mu=1}^{K} p_\mu \ln \left(\frac{p_\mu}{q_\mu} \right). \tag{770}$$

You should recognize the second term as the Kullback–Leibler divergence between the probability distributions $\mathbf{p} \equiv \{p_\mu\}$ and $\mathbf{q} \equiv \{q_\mu\}$, from Eq (656).

$$D_{\text{KL}}(\mathbf{p}||\mathbf{q}) \equiv \sum_{\mu=1}^{K} p_\mu \ln \left(\frac{p_\mu}{q_\mu} \right). \tag{771}$$

We recall that the KL divergence measures the cost of coding signals with the wrong distribution. Equation (770) shows us that better knowledge of the probability distribution doesn't just allow us to make shorter codes. The amount by which we can compress the data describing the sequence of winners in the horse race is exactly the amount by which our winnings can grow. More precisely, if we can build a shorter code than the one built implicitly by the track operators, then we will gain exactly in proportion to this shortening. Thus, in this context, we literally get paid for constructing more efficient representations of the data (!).

We have connected the growth rate of winnings to the efficiency with we can represent data, but this isn't quite as compelling as a direct connection to how much information we have about the outcome of the game, which

is where we started in the case of coin flips; let's see if we can do better. Imagine that, on each trial i, we have access to some signal x_i that tells us something about the likely outcome. More precisely, when we observe x_i, the probability that the outcome will be μ on trial i is not p_μ but rather some conditional probability $p(\mu|x_i)$; if the signals x are themselves chosen from some distribution $P(x)$, then for consistency we must have

$$p_\mu = \int dx\, P(x) p(\mu|x). \tag{772}$$

To use the extra information provided by the signal x, you will adjust your strategy to bet a fraction $f_\mu(x_i)$ on the outcome μ given that you have 'heard' x_i. How does the extra information provided by x improve your winnings?

To compute the growth of winnings in the presence of extra information, we proceed along the same lines as before, to find the analog of Eq (762):

$$\Lambda[\{f_\mu(x)\}] = \int dx P(x) \sum_{\mu=1}^{K} p(\mu|x) \ln[f_\mu(x) g_\mu]. \tag{773}$$

Now we need to maximize this, choosing strategies that are defined by the *functions* $f_\mu(x)$, where for each x we have the constraint that $\sum_\mu f_\mu(x) = 1$. Once again the solution to this optimization problem is proportional gambling, but now the proportions are conditioned on your knowledge, so that the analog of Eq (767) becomes

$$f_\mu^{\rm opt}(x) = p(\mu|x). \tag{774}$$

This determines the optimal growth rate,

$$\Lambda_{\rm opt} = \int dx P(x) \sum_{\mu=1}^{K} p(\mu|x) \ln[p(\mu|x) g_\mu]. \tag{775}$$

Problem 146: Fill in the steps leading to the derivation of $\Lambda[\{f_\mu(x)\}]$ in Eq (773) and the consequences of optimizing this functional, Eq's (774) and (775).

The important result is the gain in growth rate that is possible by virtue of having access to the signal x, that is the difference between $\Lambda_{\rm opt}$ in Eq (775) and Eq (768):

$$\begin{aligned}
\Delta\Lambda_{\rm opt} &= \int dx P(x) \sum_{\mu=1}^{K} p(\mu|x) \ln[p(\mu|x) g_\mu] \\
&\quad - \sum_{\mu=1}^{K} p_\mu \ln[p_\mu g_\mu] \qquad (776) \\
&= \int dx P(x) \sum_{\mu=1}^{K} p(\mu|x) \ln[p(\mu|x) g_\mu] \\
&\quad - \int dx P(x) \sum_{\mu=1}^{K} p(\mu|x) \ln[p_\mu g_\mu] \qquad (777) \\
&= \int dx P(x) \sum_{\mu=1}^{K} p(\mu|x) \ln\left[\frac{p(\mu|x)}{p_\mu}\right]. \qquad (778)
\end{aligned}$$

We see that the details of the payoffs g_μ drop out, and that *the gain in growth rate is exactly the mutual information between the signal x and the outcomes μ.*

Once again information translates directly into the (increased) rate at which capital can grow. Thus, the abstract measure of information has a clear impact on very down to earth measures of performance in a real world task. But, beyond metaphor,[83] what does this have to do with life?

The most direct connection between life and gambling is through the phenomenon of persistence. Many bacteria have two distinct lifestyles. In one (for example), they grow quickly in most environments, but are very susceptible to being killed by antibiotics. In the other, they grow very slowly, but survive the antibiotics. This is almost exactly the horse race—if the bacterium bets correctly, it grows, but if it bets incorrectly it dies (or grows at rates far below what is possible). Absent any direct measurements on the environment, a population of genetically identical bacteria will maximize its growth rate by a form of proportional gambling, so that even in a healthy person, not taking antibiotics, we should see that some of the resident bacteria persist in a state of slow growth and (eventual) antibiotic resistance;[84] the fraction of bacteria in this states reflects the population's estimate of the probability that they will encounter the hostile environment of antibiotics [do we know anything about whether bacteria are doing this correctly?]. We also see that gaining information about the environment opens the possibility of faster growth, in precise proportion to the information gained.

---

[83] Life is a gamble, etc..

[84] Here "resistance" is used colloquially. Technically, antibiotic resistance refers to a trait which is encoded genetically, and hence inheritable, rather than a lifestyle choice. The (choosable) state in which bacteria grow slowly but are not killed by antibiotics is called "persistent."

In a world of two alternatives, there is not much information to gain. There are examples of bacteria that choose among a wider variety of lifestyles, and these phenomena (including the simple example of two alternatives) are called 'phenotypic switching.' In the approximation that for each environment there is only one phenotype which grows, phenotypic switching is exactly the horse racing problem.

Problem 147: Something based on phenotypic switching .. look through Kussell et al for ideas.

The example of phenotypic switching makes a nice map back to the early work about gambling, but is perhaps still a bit too simple. Let's try to be more general. Imagine a bacterium that lives in an environment in which the concentrations of nutrients are fluctuating (slowly, so we don't have to worry about dynamics). In order to make use of the currently available nutrients, the bacterium must express the relevant enzymes involved in metabolism. Let's simplify and assume that there is one nutrient or substrate at concentration s and one relevant gene at expression level g. The bacterium will then grow at some rate $r(s,g)$ that depends both on the state of the world (s) and on its internal state (g).

The growth rate of the bacterium is a compromise between two effects. On the one hand, growth requires metabolism of the available nutrient, and so growth should be faster if there is either more nutrient or more enzyme. On the other hand, making the enzyme itself takes resources, and this should slow the growth; in the limit of small nutrient concentrations, this cost can become dominant, and growth would stop if the cell tried to make too much enzyme. This scenario is shown schematically in Fig 140.

Problem 148: A simple fitness landscape. The schematic in Fig 140 is based on a simple model. Suppose that growth is precisely proportional to the rate at which the enzyme degrades the substrate. In a Michaelis–Menten kinetic scheme for the enzyme [poiner to earlier discussion of MM kinetics], this means that the rate of degradation (in molecules per second) will be

$$V = V_{\max} g \frac{s_{\rm free}}{K + s_{\rm free}}, \tag{779}$$

where g is the number of copies of the enzyme molecule, $V_{\max}$ is the maximum rate at which the enzyme can run, $s_{\rm free}$ is the concentration of the substrate free in solution, and K is the 'Michaelis constant' that sets the scale for half–saturation of the enzyme. The total substrate concentration is the sum of that free in solution and bound to the enzyme,

$$s = s_{\rm free} + \frac{1}{\Omega} g \frac{s_{\rm free}}{K + s_{\rm free}}, \tag{780}$$

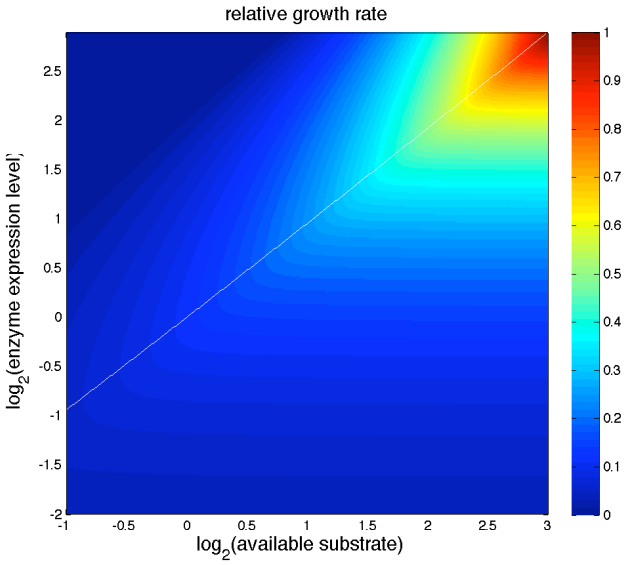

FIG. 140 A schematic of bacterial growth rate as a function of available substrate concentration and enzyme expression level. The growth rate is a compromise between metabolizing the substrate and the cost of making the enzyme. The thin white line [redraw!] traces the optimal setting of expression level as a function of substrate availability.

where Ω is the cell volume. If the growth rate is proportional to the metabolic rate, less a correction for the cost of making the enzymes, we should have

$$r(s,g) = \alpha g \frac{s_{\rm free}}{K + s_{\rm free}} - \beta g. \tag{781}$$

Solve for $s_{\rm free}$ to rewrite $r(s,g)$ explicitly in terms of s. Then show that by proper choice of units, there is only one arbitrary parameter. What is the meaning of this remaining parameter? Make some reasonable choices, and plot your own version of Fig 140.

Imagine a bacterium whose life is governed by Fig 140. As the available substrate concentration fluctuates, one possibility is that all bacteria carefully adjust their enzyme expression levels to achieve optimal growth rate under each condition. An extreme alternative is that different bacteria in the population choose their expression levels at random out of some distribution, and hope that some of them by chance have made good choices, much as in the proportional gambling scenario. In the first case, the expression level carries an enormous amount of information about the concentration of available substrate—indeed, if we imagine that the optimum is traced perfectly, then knowing the expression level would tell us the exact substrate concentration, and this represents an infinite amount of information (!). In contrast, the gambling strategy involves no correlation of the internal and external states, and hence no information in conveyed.

Evidently, the average growth rate across an ensemble of environments will be larger if the bacteria can adjust their expression levels perfectly, but maybe this is so obvious as not to be interesting. We know that there is some average growth rate which can be achieved with no information about the outside world, and that an infinite amount of information would allow the population to grow faster. What happens in between?

The mutual information between the internal state g and the external world s can be written as

$$I(g;s) = \int ds\, P(s) \int dg\, P(g|s) \log_2 \left[\frac{P(g|s)}{P(g)} \right]. \quad (782)$$

We can make $I(g;s)$ as small as we like by letting $P(g|s)$ approach $P(g)$. But suppose that we want to maintain some average growth rate in the ensemble of environments defined by $P(s)$. This average growth rate is

$$\langle r \rangle = \int ds\, P(s) \int dg\, P(g|s) r(s,g). \quad (783)$$

Now it seems clear that not all conditional distributions $P(g|s)$ are consistent with a given $\langle r \rangle$. What we would like to show is that there is a minimum value of $I(g;s)$ consistent with $\langle r \rangle$.

The problem we have is a constrained minimization, so as usual we introduce a Lagrange multiplier and minimize

$$\mathcal{F}[P(g|s)] \equiv I(g;s) - \lambda \langle r \rangle - \int ds\, \mu(s) \int dg\, P(g|s), \quad (784)$$

where the second set of Lagrange multipliers $\mu(s)$ enforces normalization of the distributions $P(g|s)$ at each value of s. Finding the minimum in this case is straightforward. The key step is to evaluate the derivative of the information with respect to the conditional distribution:

$$\frac{\delta I(g;s)}{\delta P(g|s)} = \frac{\delta}{\delta P(g|s)} \int ds\, P(s) \int dg\, P(g|s) \log_2 \left[\frac{P(g|s)}{P(g)} \right] \quad (785)$$

$$= P(s) \log_2 \left[\frac{P(g|s)}{P(g)} \right] + \frac{1}{\ln 2} P(s) P(g|s) \cdot \frac{1}{P(g|s)} - \frac{1}{\ln 2} \int ds'\, P(s') P(g|s') \frac{1}{P(g)} \frac{\delta P(g)}{\delta P(g|s)} \quad (786)$$

$$= P(s) \log_2 \left[\frac{P(g|s)}{P(g)} \right] + \frac{1}{\ln 2} P(s) - \frac{1}{\ln 2} P(g) \frac{1}{P(g)} P(s) \quad (787)$$

$$= P(s) \log_2 \left[\frac{P(g|s)}{P(g)} \right], \quad (788)$$

which is nice because all the messy bits cancel out. Now we can solve our full problem:

$$0 = \frac{\delta \mathcal{F}[P(g|s)]}{\delta P(g|s)} \quad (789)$$

$$= \frac{\delta}{\delta P(g|s)} \left[I(g;s) - \lambda \int ds\, P(s) \int dg\, P(g|s) r(s,g) - \int ds\, \mu(s) \int dg\, P(g|s) \right] \quad (790)$$

$$= P(s) \log_2 \left[\frac{P(g|s)}{P(g)} \right] - \lambda P(s) r(s,g) - \mu(s) \quad (791)$$

$$\log_2 \left[\frac{P(g|s)}{P(g)} \right] = \lambda r(s,g) + \frac{\mu(s)}{P(s)} \quad (792)$$

$$P(g|s) = \frac{1}{Z(s)} P(g) \exp\left[\beta r(s,g) \right], \quad (793)$$

where $\beta = \lambda \ln 2$, and $Z(s) = \exp[\ln 2 \mu(s)/P(s)]$ is a normalization constant,

$$Z(s) = \int dg\, P(g) \exp\left[\beta r(s,g) \right], \quad (794)$$

and of course we must obey

$$P(g) = \int ds\, P(s) P(g|s). \quad (795)$$

Notice that our solution for $P(g|s)$ is (roughly) a Boltzmann distribution, where $-r(g,s)$ plays the role of the energy and β is the inverse temperature. As expected

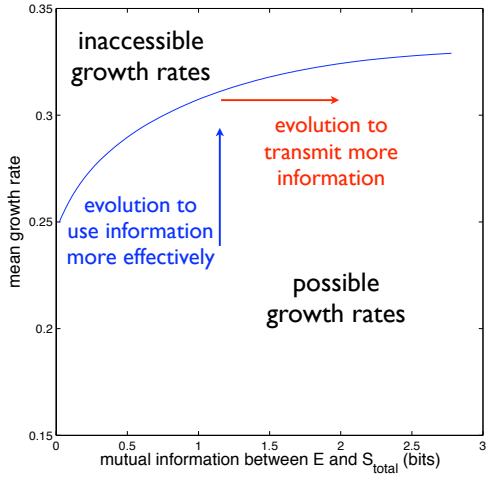

FIG. 141 Mean growth rate as a function of the mutual information between expression levels and substrate availability for the system in Fig 140. We assume that the (log) substrate is chosen from a distribution that is uniform over the 16–fold range shown in Fig 140, and then we solve for the optimal $P(g|s)$ using Eq's (793–794).

from this analogy, we can write the information and average growth rate as derivatives,

$$I(g;s) = \lambda\langle r\rangle - \int ds\, P(s)\log_2 Z(s), \tag{796}$$

$$\langle r\rangle = \int ds\, P(s)\frac{d\ln Z(s)}{d\beta}. \tag{797}$$

The Boltzmann form of the optimal solution in Eq (793) helps our intuition. At small β, the distribution $P(g|s)$ is almost the same as $P(g)$, so that very little information is conveyed between internal and external states. In contrast, as $\tilde{\lambda}$ becomes large, the distribution $P(g|s)$ becomes sharply peaked around the value expression level $g_{\rm opt}(s)$ that maximizes the growth rate. Varying β should trace out a curve of mean growth rate vs. information, and this is shown in Fig 141. We see from the derivation that this curve represents the maximum mean growth rate achievable given a certain amount of mutual information, or alternatively the minimum amount of information required to achieve a certain mean growth rate, $I_{\rm min}(\langle r\rangle)$.

Problem 149: Asymptotics of growth rate vs information. The precise form of the relationship between the mean growth rate and the minimum information depends, of course, on details of the function $r(s,g)$. Show that the behavior at large values of the minimum information is more nearly universal. To do this, develop an asymptotic expansion at large values of λ,

$$P(g|s) = \frac{1}{Z(s)}P(g)\exp\left[\tilde{\lambda}r(s,g)\right]$$
$$\approx \frac{1}{Z(s)}P(g)\exp\left[\tilde{\lambda}r(s,g_{\rm opt}(s)) + \frac{\tilde{\lambda}}{2}A(g-g_{\rm opt}(s))^2\right], \tag{798}$$

$$A = \left.\frac{\partial^2 r(s,g)}{\partial g^2}\right|_{g=g_{\rm opt}(s)} \tag{799}$$

and use this expansion to evaluate $Z(s)$, from which you can calculate $I_{\rm min}(\langle r\rangle)$. Can you generalize your discussion to the case where there are many substrates and many genes to control?

It is important to take seriously the scales in Fig 141. It could have been that the full growth advantage derived from controlling expression levels was achievable with only a small fraction of a bit, or conversely that it required many tens of bits. In fact, for this simple problem the answer is that cells can make use of more than one bit, but not too much more. This means that (near–)optimal growth requires more than just turning a gene on and off, and presumably this is even more clear if we think about more realistic situations where there are multiple substrates and multiple genes. As we will see in the next section, the noise levels measured for the control of gene expression set a limit of $\sim 1-3$ bits to the information that can be transmitted through these control elements. Thus, the amount of information that cells need in order to optimize their growth in varying environments is plausibly close to the maximum they can transmit, and this limit in turn is set by the number of molecules that the cell is devoting this these tasks.

Just to be clear, it's useful to think about the alternatives. If information is cheap, so that it is easy for cells to transmit many bits, then evolution selects for mechanisms that drive the system upward in the information/fitness plane of Fig 141. But if information itself is hard to come by, evolutionary pressure (which really only acts to increase growth rates) must necessarily drive cells outward along the information axis.

Sometimes the fact that organisms have to be flexible and survive in a fluctuating environment is offered as a qualitative argument against the possibility of optimization. Indeed, if the environment fluctuates, it may not be advantageous for organisms to drive toward "perfect" performance under any one set of conditions. But the argument we have given here shows that strategies for dealing with varied environments are themselves subject to optimization, making the most of a limited amount of information and eventually being pushed by selection to gather more bits.

In the problem of horse races, or phenotypic switching, information translated directly into a growth rate. Here

we see that, more generally, there is a minimum amount of information needed to achieve a given average growth rate. In both of these cases, information is necessary and permissive, but not sufficient. Thus organisms *can* grow faster if they gather and represent more information, but this is not guaranteed—they might make poor use of the information, and fail to reach the bound on their growth rate. We have focused here on achieving a certain average growth rate, but it should be clear that the whole discussion can be transposed to other domains. For example, if I ask you to point at a target that can appear at random in your visual field, and reward you in proportion to how close you come to the exact position of the target, then in order to collect a certain level of average reward your brain must represent some minimum amount of information about the target location. Quite generally, we can imagine plotting some "biological" measure of performance—probability of catching a mate, nutritional value extracted from picking fruit, growth rate, happiness, ... —versus the amount of information that the organism has about the relevant variables. This "information/fitness" plane will be divided by a curve which separates the possible from the impossible, since without a certain minimum level of information, higher fitness is impossible.

Problem 150: Information and motor control. Give a simple example, maybe from smooth pursuit?

In the information theory literature, the sort of bounds we are computing here go by the name of "rate–distortion" curves. For example, if we measure image quality by some complicated perceptual metric, then to have images of a certain quality, on average, we will need to transmit a minimum number of bits. In this spirit, we can think about more complicated situations, such as organisms foraging or acting in response to sensory stimuli and collecting rewards. Although one is not rewarded specifically for bits, the message of rate–distortion theory is that to collect rewards at some desired rate will always require a minimum number of bits of information.

In constructing a rate distortion curve, we implicitly define some bits as being more relevant than others. Thus if I need to match my state to that of the environment, presumably some environmental variables need to be tracked more accurately than others; since the rate distortion curves gives the minimum number of bits, I need to get this right and put the precision (extra bits) in the right place. This is important, because it means that we have a framework for assigning value to bits. To be concrete, in Fig 794 it is possible to imagine an infinite variety of mechanisms that gather the same number of bits but fail to achieve the maximum mean growth rate, either because the use the bits incorrectly or because they have gathered the wrong bits. Bits in and of themselves are not guaranteed to be useful, but to do useful things there is a minimum number of bits that we need.

An interesting if unfinished connection of rate–distortion theory to biological systems is the case of protein structure. If I want to describe protein structures with high precision, I need to tell you where every atom is located. But if sequence determines structure, then to some accuracy I just need to tell you the amino acid sequence, which is at most $\log_2(20)$ bits per amino acid, and many fewer per atom. In fact, as we have discussed in Section III.A, many different sequences generate essentially the same structure, so there must be an even shorter description. Thus, if we imagine taking the ensemble of real protein structures, there must be a description in very few bits that nonetheless generates rather small errors in predicting the positions of the atoms. Finding the optimally compact description (i.e., along the true rate–distortion curve) would be a huge help in understanding protein folding, because the joint table of sequences and (compactly described) structures would be much smaller. There is even an intuition that there must be such a compact representation with high accuracy in order to make folding rapid, essentially because the number of states needed for an accurate description should be connected to the number of states that the protein much "search" through as it folds. I am not sure how to make this rigorous, but it's interesting.

Problem 151: Clustering structures. Give an example of constructing rate–distortion curves via clustering ... maybe something plausibly connected to molecular structures?

We can search for compact descriptions of protein structure by approximating the local path of the $\alpha-$carbon backbone as moves on a discrete lattice, making the lattic progressive more complex. We can do better by moving off the lattice to cluster the natural dihedral angles describing the path from one amino acid to the next Be sure we talk about ϕ, ψ description of proteins before this, and point back]; results are shown in Fig 142. Indeed, by the time we have assigned 10 or 20 states per amino acid, we can reconstruct structures with $1-2$ Å rms accuracy.

Another very specific connection between biology and bits is in the case of embryonic development. In the simplest model of morphogen gradients, each independently "reads out" the local concentration of the morphogen(s),

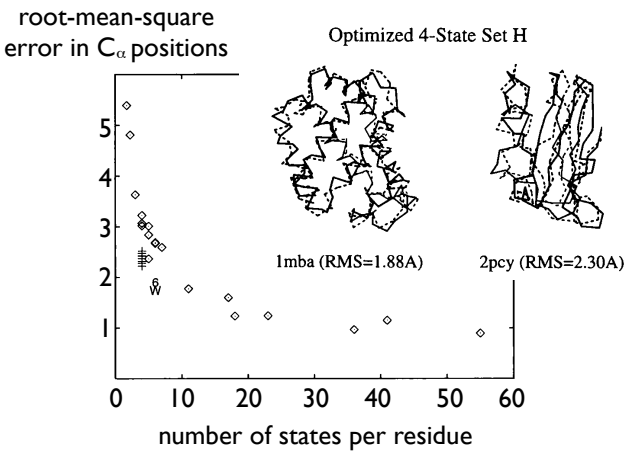

FIG. 142 Rate–distortion curve (or its moral equivalent) for protein structures, from Park and Levitt (1995). The path of the $\alpha-$carbon backbone is approximated by a discrete set of local 'moves' along the chain, which is roughly equivalent to forcing the structure to live on a lattice. Diamonds correspond to lattices with different structures (e.g., 3 possible moves on a tetrahedral lattice); + and W correspond to discrete approximations obtained by clustering known structures based on the Ramachandran angles at each site. Plotted on the y−axis is the root mean square error in the positions of all the $\alpha-$carbons along the chain. Inset shows two examples of protein structures, compared with their discrete approximations.

and makes decisions—most importantly, about the regulation of gene expression—based on this local measurement, as in Fig 143. In this model, the only thing that a cell knows about its position in the embryo is the morphogen concentration, and so the information that cells have about position can be no larger than the information that they extract about this concentration. In effect there is a communication channel from the morphogen to the expression levels of the genes which defines the blueprint for development, and the information that can be transmitted along this channel sets a bound on the complexity and reliability of the blueprint. As an example, if we have N rows of cell along one axis of the embryo, and each row reliably adopts a distinct fate that we can 'see' by looking at the expression levels of a handful of genes, then (again, in the simplest model) there must be $\log_2 N$ bits of information transmitted through the regulatory network that takes the morphogens as input and gives the gene expression levels as output. As in the discussion of growth rates, this becomes interesting because, as we shall see, the information capacity of gene regulatory elements is quite limited. Rough estimates of the relevant quantities in the *Drosophila* embryo suggest that the embryo might indeed be forming patterns near the limits set by the information capacity of gene regulation.

What happens if things are more complicated than in Fig 143? In particular, we know about plenty of systems which form patterns spontaneously, without any analog of the "maternal" signal to break the translational symmetry. It is important to realize that while patterns can form spontaneously, information can't really be created, only transmitted. In a crystal, for example, once we know that one atom is in a particular position we can predict the position of other atoms, but this is only because of the bonds that connect the atoms. Because all the atoms undergo Brownian motion, the transmission of information is not perfect, and knowledge of one atomic position provides only a limited number of bits about the position of another atom; this limit on information transmission becomes tighter as the temperature—and hence the noise level in the "communication channel" which connects the distant atoms—becomes larger, until the crystal melts and there is no information transmitted over long distances.

Problem 152: Transmitting positional information in a crystal. Take the students through an explicit calculation of the mutual information between positions of atoms in a harmonic solid.

In non–equilibrium systems, such as the Rayleigh–Bernard convection cell shown in Fig 144, we see spatial patterns in which some local variable such as the temperature, fluid density or velocity at one position predicts

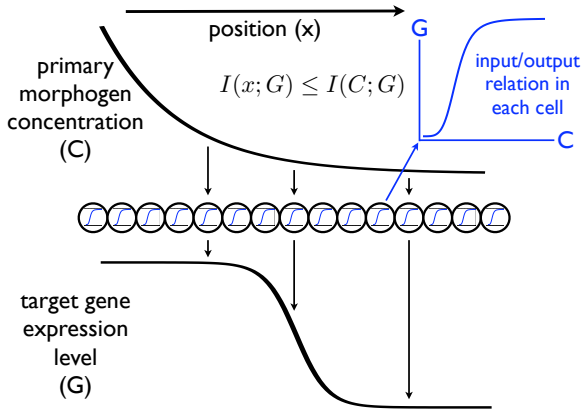

FIG. 143 Information flow in a "feed–forward" model of genetic control in the early embryo. The concentration C of the primary morphogen depends on position x, and each cell responds independently by modulating the expression level G of some target gene (or genes). In this simple view, information about position only reaches the gene expression level through the intermediary of the primary morphogen concentration, and hence we have $I(x;G) \leq I(C;G)$.

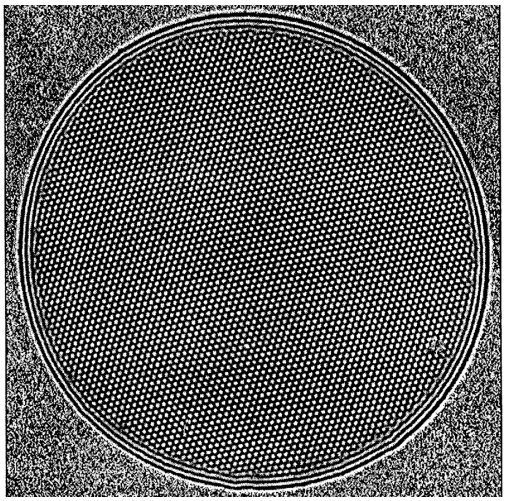

FIG. 144 This looks like a perfect crystal of beads, but it actually is a small ($\sim$ 10 cm diameter) container filled with carbon dioxide at high pressure, and heated from below. The image is formed by passing light through the gas, sometimes called a 'shadowgraph.' The temperatures at the top and bottom of the container are held very constant (to within a few thousandths of a degree) so that the patterns will not be disrupted by variations in conditions; similarly, the top and bottom of the container are extremely flat (smooth to within the wavelength of light), and the whole system is held horizontal with high precision so that the direction of gravity is aligned with the axis of symmetry through the center of the circle. From E Bodenschatz et al (1991).

the value of the corresponding variable at another position. If we call this local variable $\phi(x)$, then if we imagine a large ensemble of snapshots like the one in Fig 144, we can build up the distribution functional $P[\phi(\vec{x})]$. The statement that we have a periodic pattern, for example, is the statement that if we look at two points separated by an appropriately chosen vector $\vec{d}$, then $\phi(\vec{x}) \approx \phi(\vec{x}+\vec{d})$. But if we point to the first point $\vec{x}$ at random, we can get a broad range of values for $\phi_1 \equiv \phi(\vec{x})$, drawn from a distribution $P_1(\phi_1)$. Similarly, if we are choosing $\vec{x}$ at random then $\phi_2 \equiv \phi(\vec{x}+\vec{d})$ is also broadly distributed; in fact, it must come from the same distribution as ϕ_1. But once we know ϕ_1, if there is a periodic pattern then the distribution $P(\phi_2|\phi_1)$ must be sharply peaked around $\phi_1 = \phi_2$, and hence very different from the "prior" distribution of ϕ_2. But this is exactly the condition for there to be mutual information between ϕ_1 and ϕ_2. Thus, the existence of a spatial pattern is equivalent to the presence of mutual information between the local variables at distant points. Where does this information come from? As with the bonds connecting the atoms in the crystal, it must be transmitted through the dynamics of the system, which connect points only to their immediate neighbors.

In a strict interpretation of the concept of positional information in embryo, we actually require more than mutual information between local variables at distant points. We require that the value of some local variable(s), typically the expression levels of several genes, tell us about the location of the point where we have observed them. In this way, cells would "know" their position in the embryo by virtue of their expression levels, and these signals could drive further processes in a way that is appropriate to the cell's location—not just relative to other cells, but in absolute terms.[85] If we call the local variables $\{g_i\}$, for gene expression levels, then the positional information is $I(\vec{x};\{g_i\})$. But the local variables at point x are controlled by a set of inputs which may include external, maternally supplied morphogens, the expression levels $\{g_i\}$ in neighboring cells, and perhaps other variables as well. We can always write the distribution of expression levels at one point in terms of this inputs,

$$P(\{g_i\}|x) = \int d(\text{inputs})\, P(\{g_i\}|\text{inputs})P(\text{inputs}|x). \tag{800}$$

Noise in the control of gene expression corresponds to the fact that the distribution $P(\{g_i\}|\text{inputs})$ is not infinitely narrow. Now because, at any one point, information flows from x to the inputs to the $\{g_i\}$, we must have $I(x;\{g_i\}) \leq I(\text{inputs};\{g_i\})$, and this is true no matter how complicated the inputs might be. More importantly, as hinted at in the analysis of the first synapse in fly vision (Fig 139), any input/output device has a maximum amount of information it can transmit that is determined by its noise level. Thus, if we think of all the whole network of interactions that result in the regulation of the gene expression levels $\{g_i\}$, the noise in this network determines a maximum value for $I(\text{inputs};\{g_i\})$, and this sets a limit to the amount of positional information that cells in the embryo can acquire and encode with these genes.

Problem 153: The data processing inequality. What we need in the previous paragraph is a special case of a more general inequality .. derive it.

To summarize, the reliability and complexity of the patterns that can form during embryonic development

---

[85] This is certainly what "positional information" means in the usual descriptions of the concept; see the discussion of the information carried by Hunchback expression levels in the fly embryo, surrounding Fig 136. There are almost no measurements of this information, in bits, so it remains possible that real cells know much more about their relative position than about their absolute position. This wouldn't change the spirit of what I am saying here, but the details would matter. This is one of many open questions about information flow in the embryo.

are limited by the amount of positional information that cells can acquire and represent. This information in turn is limited by the "capacity" of the genetic or biochemical networks whose outputs encode the positional information. Therefore, if real networks operate in a regime where this capacity is small, the complexity of body plans will be limited by the ability of the organism to squeeze as much information as possible out of these systems.

Most of the examples we have considered thus far have the feature that the information is "about" something that has obvious relevance for the organism. Can we find some more general way at arriving at such notions of relevant? It is useful to have in mind an organism collecting a stream of data, whether the organism is like us, with eyes and ear, or like a bacterium, sensing the concentrations of various molecules in its external and internal environment. Of all these data, the only part we can use to guide our actions (and eventually collect rewards, reproduce, etc.) is the part that has predictive power, since by the time we act we are already in the future. Thus we can ask how to squeeze, out of all the bits we collect, only those bits which are relevant for prediction.

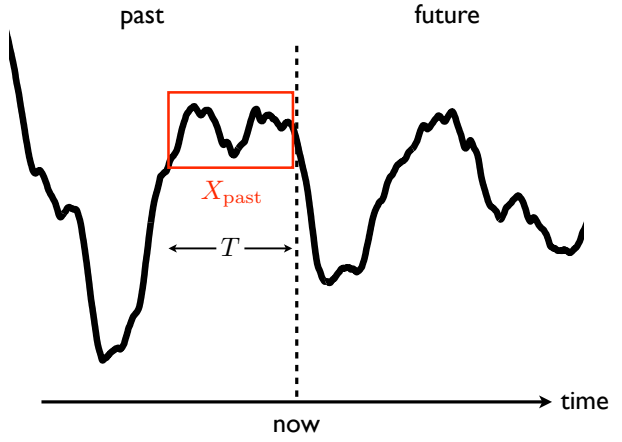

FIG. 145 A schematic of the prediction problem. We observe a time series, and at some moment (now) we look back at a segment of the recent past with duration T, $X_{\rm past}$. From this, we try to infer something about what will evolve in the future.

[I am worried that this goes a little quickly.] More concretely, as in Fig 145, if we observe a time series through a window of duration T (that is, for times $-T < t \leq 0$), then to represent the data $X_{\rm past}$ we have collected requires $S(T)$ bits, where S is the entropy, but the information that these data provide about the future $X_{\rm future}$ (i.e., at times $t > 0$) is given by some $I(X_{\rm past}; X_{\rm future}) \equiv I_{\rm pred}(T) \ll S(T)$. In particular, while for large T the entropy $S(T)$ is expected to become extensive, the predictive information $I_{\rm pred}(T)$ always is subextensive. Thus we expect that the data $X_{\rm past}$ can be compressed significantly into some internal representation $X_{\rm int}$ without losing too much of the relevant information about $X_{\rm future}$. Formally, we can construct the optimal version of this mapping by solving

$$\max_{X_{\rm past}\to X_{\rm int}} \left[I(X_{\rm int}; X_{\rm future}) - \lambda I(X_{\rm int}; X_{\rm past}) \right], \quad (801)$$

where $X_{\rm past} \to X_{\rm int}$ is the rule for creating the internal representation and λ is a Lagrange multiplier. This sort of problem has been dubbed an 'information bottleneck,' because we try to preserve the relevant information while squeezing the input data through a narrow channel.

Problem 154: Predictive information is subextensive. If we observe a stationary stochastic process, $x(t)$, on the interval $t_1 < t \leq t_1 + T$, the entropy of the distribution $P[x(t)]$ depends only on T, not t_1; let's call this entropy $S(T)$.

(a.) Use your intuition from statistical mechanics to explain why we expect $S(T)$ to grow extensively, that is $S(T) \propto T$ at large T. More formally, show that at large T

$$S(T) \to \mathcal{S}T + S_1(T), \quad (802)$$

where

$$\lim_{T\to\infty} \frac{S_1(T)}{T} = 0. \quad (803)$$

Thus, although $S_1(T)$ can grow with T, it must grow more slowly than T itself—it is "subextensive."

(b.) Consider the case where time is discrete, and x is Markovian, so that $x(t+1)$ depends on $x(t)$, but no earlier history. Show that, in this case, $S_1(T)$ is just a constant.

(c.) Consider the case where $X_{\rm past} \equiv x(-T < t \leq 0)$ and $X_{\rm future} \equiv x(0 < t < T')$. Show how the predictive information $I_{\rm pred}(T, T') \equiv I(X_{\rm past}; X_{\rm future})$ is related the function $S(T)$; you should be able to do this in general, without the Markov assumption. Show further that there a finite limit as the duration of the future becomes infinite, and that this limit $I_{\rm pred}(T)$ is subextensive.

In general, we should consider the mapping $X_{\rm past} \to X_{\rm int}$ to be probabilistic, so we can describe it by some conditional distribution $P(X_{\rm int}|X_{\rm past})$. Then the quantity we are trying to maximize becomes

$$-\mathcal{F} = \sum_{X_{\rm int},X_{\rm past}} P(X_{\rm int}|X_{\rm past})P(X_{\rm past}) \log_2 \left[\frac{P(X_{\rm int}|X_{\rm past})}{P(X_{\rm int})}\right] - \lambda \sum_{X_{\rm int},X_{\rm future}} P(X_{\rm int}|X_{\rm future})P(X_{\rm future}) \log_2 \left[\frac{P(X_{\rm int}|X_{\rm future})}{P(X_{\rm int})}\right]. \quad (804)$$

This is written as if our choice of representation $X_{\rm int}$ depends directly on the future, but of course this isn't true; any correlation between what we write down and what happens in the future is inherited from the data that we collected in the past,

$$P(X_{\rm int}|X_{\rm future}) = \sum_{X_{\rm past}} P(X_{\rm int}|X_{\rm past})P(X_{\rm past}|X_{\rm future}). \quad (805)$$

In addition, we have

$$P(X_{\rm int}) = \sum_{X_{\rm past}} P(X_{\rm int}|X_{\rm past})P(X_{\rm past}). \quad (806)$$

As usual, we have to take the derivative of $\mathcal{F}$ with respect to the distribution $P(X_{\rm int}|X_{\rm past})$, being careful to add a Lagrange multiplier $\mu(X_{\rm past})$ that fixes the normalization for each value of $X_{\rm past}$), and then we set the derivative to zero to find an extremum. Since the optimization of $\mathcal{F}$ is independent of multiplicative factors, we can make things simpler by taking natural logs instead of logs base 2. Then the algebra is as follows:

$$0 = \frac{\delta}{\delta P(X_{\rm int}|X_{\rm past})} \left[-\mathcal{F} - \sum_{X_{\rm past}} \mu(X_{\rm past}) \sum_{X_{\rm int}} P(X_{\rm int}|X_{\rm past}) \right] \quad (807)$$

$$= P(X_{\rm past}) \ln \left[\frac{P(X_{\rm int}|X_{\rm past})}{P(X_{\rm int})}\right] - \lambda \sum_{X_{\rm future}} P(X_{\rm past}|X_{\rm future})P(X_{\rm future}) \ln \left[\frac{P(X_{\rm int}|X_{\rm future})}{P(X_{\rm int})}\right] - \mu(X_{\rm past}). \quad (808)$$

To proceed, it would be useful to divide through by a factor of $P(X_{\rm past})$, at which point we have

$$\ln \left[\frac{P(X_{\rm int}|X_{\rm past})}{P(X_{\rm int})}\right] = \lambda \sum_{X_{\rm future}} P(X_{\rm future}|X_{\rm past}) \ln \left[\frac{P(X_{\rm int}|X_{\rm future})}{P(X_{\rm int})}\right] + \tilde{\mu}(X_{\rm past}), \quad (809)$$

where $\tilde{\mu}(X_{\rm past}) = \mu(X_{\rm past})/P(X_{\rm past})$. Further, since on the right we have a conditional distribution of $X_{\rm future}$ given $X_{\rm past}$, it would be nice to rearrange the ratio inside the logarithm,

$$\frac{P(X_{\rm int}|X_{\rm future})}{P(X_{\rm int})} = \frac{P(X_{\rm future}|X_{\rm int})}{P(X_{\rm future})} = \frac{P(X_{\rm future}|X_{\rm int})}{P(X_{\rm future}|X_{\rm past})} \cdot \frac{P(X_{\rm future}|X_{\rm past})}{P(X_{\rm future})}, \quad (810)$$

so that, when we substitute we find

$$\ln \left[\frac{P(X_{\rm int}|X_{\rm past})}{P(X_{\rm int})}\right] = \lambda \sum_{X_{\rm future}} P(X_{\rm future}|X_{\rm past}) \ln \left[\frac{P(X_{\rm future}|X_{\rm int})}{P(X_{\rm future}|X_{\rm past})}\right] + \lambda \sum_{X_{\rm future}} P(X_{\rm future}|X_{\rm past}) \ln \left[\frac{P(X_{\rm future}|X_{\rm past})}{P(X_{\rm future})}\right] + \tilde{\mu}(X_{\rm past}). \quad (811)$$

We recognize the first term on the right as being the (negative) Kullback–Leibler divergence between the distribution of futures given the past, and the distribution of futures given our representation $X_{\rm int}$. Further, the second term depends only on $X_{\rm past}$, and so can be absorbed into $\tilde{\mu}(X_{\rm past})$. Thus, when the dust settles, we have

$$P(X_{\rm int}|X_{\rm past}) = \frac{P(X_{\rm int})}{Z(X_{\rm past};\lambda)} \exp\Big(-\lambda D_{KL}\left[P(X_{\rm future}|X_{\rm past})||X_{\rm future}|X_{\rm int})\right] \Big), \quad (812)$$

where $Z(X_{\rm past};\lambda)$ is a normalization constant. This isn't a solution to our problem, but rather a self–consistent

equation that the solution has to satisfy. The problem we are solving is an example of selective compression, and the particular formulation of trading bits vs. bits has come to be called the "information bottleneck" problem.

Problem 155: Fill in all the details leading from Eq (807) to Eq (812).

We should think of Equation (812) as being like the result in Eq (794), but instead of adjusting an internal state in relation to the "potential" formed by the growth rate, here the effective potential is the (negative) Kullback–Leibler divergence, which measures the similarity between the distributions of futures given the actual past and given our compressed representation of the past. This means that if two past histories lead to similar distributions of futures, they should be mapped into the same value of $X_{\rm int}$. This makes sense, since we are trying to throw away any information that doesn't have predictive power. When λ is very large, differences in the expected future need to be very small before we are willing to ignore them, while at small λ it is more important that our description be compact, to we are willing to make coarser categories. As in rate–distortion theory, there is no single right answer, but rather a curve which defines the maximum amount of predictive information we can capture given that we are willing to write down a certain number of bits about the past, and along this curve there is a one parameter family of strategies for mapping our observations on the past into some internal representation $X_{\rm int}$.

Problem 156: Predictive information and optimal filtering. Imagine that we observe a Gaussian stochastic process $[x(t)]$ that consists of a correlated signal $[s(t)]$ in a background of white noise $[\eta(t)]$, that is $x(t) = s(t) + \eta(t)$, where

$$\langle s(t)s(t')\rangle = \sigma^2 \exp\left(-|t-t'|/\tau_c\right) \tag{813}$$
$$\langle \eta(t)\eta(t')\rangle = \mathcal{N}_0\delta(t-t'). \tag{814}$$

Recall (or see Section A.2) that the full probability distribution for the function $x(t)$ is

$$P[x(t)] = \frac{1}{Z}\exp\left[-\frac{1}{2}\int dt \int dt'\, x(t)K(t-t')x(t')\right], \tag{815}$$

where Z is a normalization constant.

(a.) Construct the kernel $K(\tau)$ explicitly. Be careful about the behavior near $\tau = 0$.

(b.) Break the data $x(t)$ into a past $X_{\rm past} \equiv x(t<0)$ and a future $X_{\rm future} \equiv x(t>0)$, relative to the time $t = 0$. Show that $P[x(t)]$ can be rewritten so that the only term that mixes past and future is of the form

$$\left[\int_{-\infty}^{0} dt\, g(-t)x(t)\right] \times \left[\int_{0}^{\infty} dt'\, g(t')x(t')\right], \tag{816}$$

where $g(t) = \exp(-t/\tau_0)$, with $\tau_0 = \tau_c(1+\sigma^2\tau_c/\mathcal{N}_0)^{-1/2}$. More formally, if we define

$$z = \int_{-\infty}^{0} dt\, g(-t)x(t), \tag{817}$$

show that

$$P(X_{\rm future}|X_{\rm past}) = P(X_{\rm future}|z). \tag{818}$$

Explain why the optimal internal representation of the predictive information, $X_{\rm int}$, can only depend on z.

(c.) Suppose that we are given the past data $x(t \le 0)$, and instead of being asked to predict the future, you are asked to make the best estimate of the underlying signal $s(t=0)$. [Connect back to problem in Chapter 1] Show that this optimal estimate is proportional to z.

As you just showed in the last problem, the optimal representation of predictive information is equivalent, at least in simple cases, to the separation of signals from

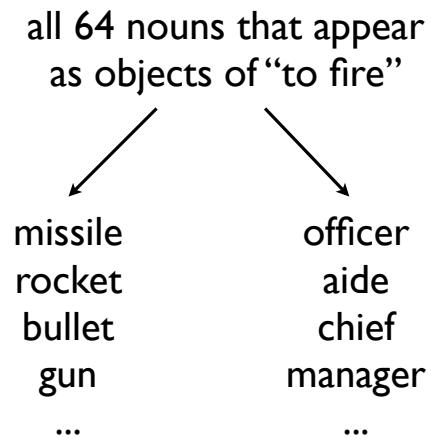

FIG. 146 A precursor of the information bottleneck problem, from Lee et al (1993). In one year of the Associated Press news reports, there are 64 nouns ($X_{\rm noun}$) which appear as the direct object of the verb "to fire," and these nouns are paired with 2147 distinct verbs ($X_{\rm verb}$). Following the ideas in the text, imagine compressing the description of the nouns, $X_{\rm noun} \to X_{\rm int}$, while trying to preserve the information that the compressed description conveys about the verb which appears with the noun. That is, maximize $I(X_{\rm int}; X_{\rm verb})$ while holding $I(X_{\rm int}; X_{\rm noun})$ fixed. Here we show the solution to the problem when $I(X_{\rm int}; X_{\rm noun}) \approx 1$ bit supports two distinct values of $X_{\rm int}$; what we list are the nouns that map to the two values of $X_{\rm int}$ with high probability. We see that this classifies the nouns by their meaning, separating weapons (firing a missile) from job titles (firing a manager). Importantly, this is based only on the co–occurrence of the nouns with verbs in sentences; there is no supervisory signal which distinguishes the different senses of the verb.

noise. In Section IV.D we will see that extracting the predictive information from other kinds of time series is equivalent to learning the underlying parameters or rules that the data obey. In a somewhat more fanciful example, we can think of $X_{\rm past}$ as a word in a sentence, and $X_{\rm future}$ as the next word; then the mapping $X_{\rm past} \to X_{\rm int}$ is equivalent to making clusters of words. When λ is small, there are very few clusters, and they correspond very closely to parts of speech. As λ becomes larger, we start to discern categories of words that seem to have meaning. Indeed the first exercise of this sort was to choose not two successive words as past and future, but rather the noun and verb in the same sentence, and then the impression (still subjective) that the resulting clusters of nouns have similar meanings is even stronger, as seen in Fig 146. It is tempting to suggest that the optimal representation of predictive information is extracting "meaning" from the statistics of sentences. [perhaps explain that some people are horrified by this suggestion?]

[Maybe say something about Tagkopolous et al (2008)? What about different ideas of predictive coding from Laughlin, Rao, ... ? Is this the place to show evidence (depending on how much we have!) that neurons provide efficient representations of predictive information, or does this go in the next section? What about a reminder that the rules of synaptic plasticity seem to know about causality, and hence might serve to build representations that favor predictive information?]

Let me try to pull the different arguments of this section together, even if imperfectly. What we really care about is how organisms can maximize some measure of performance—ultimately, their reproductive success—given access to some limited set of resources. Within any broad class of possible biological mechanisms, there is an optimum that divides the fitness/resources plane into possible and impossible regions, as in the upper right quadrant of Fig 147; evolutionary pressure drives organisms toward this boundary. But we have seen that, for any measure of fitness or adaptive value, achieving some criterion level of performance always requires some minimum number of bits; this is the content of rate–distortion theory. Thus there is a plane (in the upper left quadrant of Fig 147) of fitness vs. information, and again there is a curve that divides the possible from the impossible. Importantly, the information that an organism can use to gain a fitness advantage—even in the simple example of adjusting gene expression levels to match the availability of nutrients—is always predictive information, because the consequences of actions come after they are decided upon.

We know that bits are not free. In simple examples, such as the Gaussian channel in Section **??**, the information that can be transmitted depends on the signal to noise ratio, and this in turn depends on the resources the organism can devote, whether we are counting action potentials or molecules. If we think about the bits that will be used to direct an action, then there are many costs—the cost of acquiring the information, of representing the information, and the more obvious physical costs of carrying out the resulting actions, but we always can assign these costs to the symbols at the entrance to the communication channel. The channel capacity separates the information/resources plane into accessible and inaccessible regions, as in the lower right quadrant of Fig 147. Ideas about metabolically efficient neural codes [perhaps should be more explicit here?], for example, can be seen as efforts to calculate this curve in specific models. Of course the information we are talking about now is information that we actually collect, and this is information about the past. To close the connections among the different quantities, we need the information bottleneck, which tells us that—given the structure of the world we live in—having a certain number of bits of information about the future requires capturing some minimum number of bits about the past.

To summarize, if an organism wants to achieve a certain mean fitness, it needs a minimum number of bits of predictive power, and this requires collecting a minimum number of bits about the past, which in turn necessitates some minimum cost or available resources. Usually we

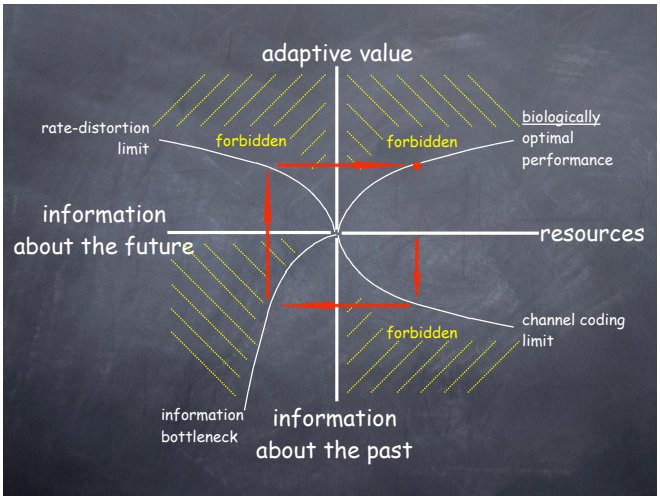

FIG. 147 Connecting the different optimization principles (Bialek et al 2007). Lines indicate curves of optimal performance, separating allowed from forbidden (hashed) regions of each quadrant. In the upper right quadrant is the biologically relevant notion of optimization, maximizing fitness or adaptive value at fixed resources. But actions that achieve a given level of adaptive value require a minimum number of bits, and since actions occur after plans these are bits about the future (upper left). On the other hand, the organism has to "pay" for bits, and hence there is a minimum resource costs for any representation of information (lower right). Finally, given some bits (necessarily obtained from observations on the past), there is some maximum number of bits of predictive power (lower left). To find a point on the biological optimum one can try to follow a path through the other three quadrants, as indicated by the arrows.

think of evolution as operating in the tradeoff between resources and fitness, but this has echoes in the other quadrants of Fig 147, where information theoretic bounds are at work. These connections provide a path whereby evolution can select for mechanisms that approach these bounds, even thought evolution itself doesn't know about bits.

The connection between information and gambling goes back to Kelly (1956). Connections of these ideas to fitness in fluctuating environments are discussed by Bergstromm & Lachmann (2005), Kussell & Leibler (2005), and more generally by Rivoire & Leibler (2011). The specific case of persistence in bacteria has been explored by Balaban et al (2004) and Kussell et al (2005); for a review see Gefen & Balaban (2009). The analogy to rate–distortion theory, demonstrating a minimum number of bits required to achieve a criterion mean growth rate, is from Taylor et al (2007); for a treatment of rate–distortion theory itself, again see Cover & Thomas (1991), in the refs to Section IV.A. Although they didn't explicitly use the language of rate–distortion theory, Park and Levitt (1995) explored the compression of protein structures into a small set of local, discrete states, asking how the complexity of this representation related to its accuracy. For a first try at connecting information flow and embryonic pattern formation, see Tkačik et al (2008). The beautiful convection patterns in Fig 144 are from Bodenscahtz et al (1991).

Balaban et al 2004: Bacterial persistence as a phenotypic switch. NQ Balaban, J Merrin, R Chait, L Kowalik & S Leibler, *Science* **305,** 1622–1625 (2004).

Bergstrom & Lachmann 2005: The fitness value of information. CT Bergstrom & M Lachmann, axXiv:q–bio.PE/0510007 (2005).

Bodenschatz et al 1991: Transitions between patterns in thermal convection. E Bodenschatz, JR de Bruyn, G Ahlers & DS Cannell, *Physical Review Letters* **67,** 3078–3081 (1991).

Gefen & Balaban 2009: The importance of being persistent: Heterogeneity of bacterial populations under antibiotic stress. O Gefen & NQ Balaban, *FEMS Microbiol Rev* **33,** 704–717 (2009).

Kelly 1956: A new interpretation of information rate. JL Kelly, Jr, *Bell Sys Tech J* **35,** 917–926 (1956).

Kussell et al 2005: Bacterial persistence: A model of survival in changing envrionments. EL Kussell, R Kishony, NQ Balaban & S Leibler, *Genetics* **169,** 1807–1814 (2005).

Kussell & Leibler 2005: Phenotypic diversity, population growth, and information in fluctuating environments. E Kussell & S Leibler, *Science* **309,** 2075–2078 (2005).

Park & Levitt 1995: The complexity and accuracy of discrete state models of protein structure. BH Park & M Levitt, *J Mol Biol* **249,** 493–507 (1995).

Rivoire & Leibler 2011: The value of information for populations in varying environments. O Rivoire & S Leibler, *J Stat Phys* **142,** 1124–1166 (2011); arXiv.org:1010.5092 [q–bio.PE] (2010).

Taylor et al 2007: Information and fitness. SF Taylor, N Tishby & W Bialek, arXiv:0712.4382 [q–bio.PE] (2007).

Tkačik et al 2008: Information flow and optimization in transcriptional regulation. G Tkačik, CG Callan Jr & W Bialek, *Proc Nat'l Acad Sci (USA)* **105,** 12265–12270 (2008).

The 'information bottleneck' was introduced by Tishby et al (1999). It has connections with statistical mechanics approaches to clustering (Rose et al 1990), and more immediate antecedents in the idea of clustering distributions of words (Pereira et al 1993). For more about predictive information, see Bialek et al (2001) and (2007). The possibility that even familiar programs of coordinated changes in bacterial gene expression may reflect (implicit) predictions is discussed by Tagkopoulos et al (2008). Energy efficiency in neural coding is discussed by Laughlin et al (1998) and by Balasubramanian et al (2001).

Balasubramanian et al 2001: Metabolically efficient information processing. V Balasubramanian, D Kimber & MJ Berry II, *Neural Comp* **13,** 799–815(2001).

Bialek et al 2001: Predictability, complexity and learning. W Bialek, I Nemenman & N Tishby, *Neural Comp* **13,** 2409–2463 (2001); arXiv:physics/0007070 (2000).

Bialek et al 2007: Efficient representation as a design principles for neural coding and computation. W Bialek, RR de Ruyter van Steveninck & N Tishby, arXiv:0712.4381 [q–bio.NC] (2007); see also *Proceedings of the International Symposium on Information Theory 2006*.

Laughlin et al 1998: The metabolic cost of neural information. SB Laughlin, RR de Ruyter van Steveninck & JC Anderson, *Nature Neurosci* **1,** 36–41 (1998).

Pereira et al 1993: Distributional clustering of English words. FC Pereira, N Tisby & L Lee, in *30th Annual Meeting of the Association for Computational Linguistics,* pp 183–190 (1993).

Rose et al 1990: Statistical mechanics and phase transitions in clustering. K Rose, E Gurewitz & GC Fox, *Phys Rev Lett* **65,** 945–948 (1990).

Tagkopoulos et al 2008: Predictive behavior within microbial genetic network. I Tagkopoulos, Y Liu & S Tavazoie, *Science* **320,** 1313–1317 (2008).

Tishby et al 1999: The information bottleneck method. N Tishby, FC Pereira & W Bialek, in *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, B Hajek & RS Sreenivas, eds, pp 368–377 (University of Illinois, 1999); physics/0004057 (2000).

C. Optimizing information flow

We have seen that organisms should care about bits—for every criterion level of performance that a system wants to achieve, there is a minimum number of bits that it needs. If bits are cheap, or easy to acquire, then this need for a minimum number of bits is true but not much of a constraint. On the other hand, if the physical constraints under which organisms operate imply severe limits on information transmission, then the minimum number of bits may approach the maximum number available, and strategies that maximize efficiency in this sense may be critical to biological function.

One of the central ideas in thinking about the efficiency with which bits can be collected and transmitted is that what we mean by efficient (and, in the extreme, optimal) depends on context, as indicated schematically in

Fig 148. In the top panel we see a typical sigmoidal input/output relation, which might describe the expression level of a gene vs. the concentration of a transcription factor, the probability of spiking in a neuron as a function of the intensity of the sensory stimulus, ... In the bottom panel we see different possibilities for the distributions out of which the input signals might be drawn. For the two distributions in blue, the input signals are confined to the saturated regions of the input/output relation, leaving the output almost always in the fully 'off' or 'on' states. In these situations, the output is always the same, and is unaffected by the changes in the input that actually occur with reasonable probability, and the system is essentially useless. More subtly, for the distribution in green, input signals are in the middle of the middle of the input/output relation, where the slope of the input/output relation is maximal, but the dynamic range of these variations is small, so that the variations in output are only a small fraction of what is possible, and these variations might well be obscured by any reasonable level of noise. Finally, for the distribution in red, the dynamic range of the likely inputs is just big enough to push the system through the full dynamic range of the input/output relation, generating large (maximal?) variations in the output.

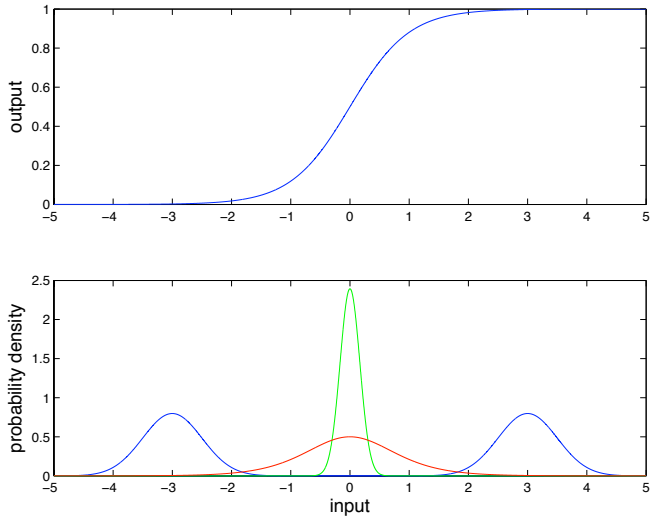

FIG. 148 At top, an example of an input/output relation. At bottom, different possible probability distributions for the inputs. As described in the text, the blue and green distributions are poorly matched to the input/output relation, while the red distribution seems to be a better match.

While it is easy for everyone to agree that, in Fig 148, the blue and green distributions of inputs are poorly matched to the input/output relation, and the red distribution is well matched, it takes a little more courage (and courts more controversy) to make a precise mathematical statement about what constitutes a good match, or the "best" match. What we will try out as a definition of "best" is that outputs should provide as much information as possible about the inputs.

Let's start with input x, chosen from a distribution $P_X(x)$, and assume that this is converted into one output y by a system that has an input/output relation $g(x)$ but also some added noise,

$$y = g(x) + \xi. \tag{819}$$

Notice that when we plot an input/output relation, as in Fig 148, we (implicitly) are referring to the *average* behavior of the system, since realistically there must be some level of noise and hence the input and output are related only probabilistically; we now make this explicit by adding the noise ξ. To keep things simple, let's assume that this noise is Gaussian, with some variance σ^2 and as usual zero mean. In principle, the variance of the output noise could depend upon the value of the input, and this will be important below, so we'll write $\sigma_y^2(x)$ to remind us that we are talking about the variance of the output (hence the subscript), but this may depend upon the input.

In order to compute the amount of information that y provides abut x, we need various probability distributions. Specifically, we want to evaluate

$$I(y;x) = \int dx \int dy\, P(x,y) \log_2 \left[\frac{P(x,y)}{P_X(x) P_Y(y)} \right] \tag{820}$$

$$= \int dx \int dy\, P(x,y) \log_2 \left[\frac{P(y|x)}{P_Y(y)} \right]. \tag{821}$$

It is the conditional distribution $P(y|x)$ that describes, in the most general setting, the probabilistic relationship between input and output. The overall distribution of outputs is given by

$$P_Y(y) = \int dx\, P(y|x) P_X(x). \tag{822}$$

With the hypothesis that the noise ξ is Gaussian, Eq (819) tells us that

$$P(y|x) = \frac{1}{\sqrt{2\pi\sigma_y^2(x)}} \exp\left[-\frac{(y - g(x))^2}{2\sigma_y^2(x)} \right]. \tag{823}$$

The information can be written (as usual) as the difference between two entropies,

$$I(y;x) = \int dx \int dy\, P(x,y) \log_2 \left[\frac{P(y|x)}{P_Y(y)} \right]$$
$$= -\int dy\, P_Y(y) \log_2 P_Y(y) - \int dx\, P_X(x) \left[-\int dy\, P(y|x) \log_2 P(y|x) \right]. \quad (824)$$

But the conditional distribution $P(y|x)$ is Gaussian, with variance $\sigma_y^2(x)$, so we can substitute for the conditional entropy from Eq (701) to give

$$I(y;x) = -\int dy\, P_Y(y) \log_2 P_Y(y) - \frac{1}{2\ln 2} \int dx\, P_X(x) \ln[2\pi e \sigma_y^2(x)]. \quad (825)$$

The distribution of outputs $P_Y(y)$ is broadened by two effects. First, as x varies, the mean value of y changes. Second, even with x fixed, noise causes variations in y. But if the noise is small, the first effect should dominate, and this will simplify our problem. Formally,

$$P_Y(y) = \int dx\, P_X(x) P(y|x) = \int dx\, P_X(x) \frac{1}{\sqrt{2\pi\sigma_y^2(x)}} \exp\left[-\frac{(y-g(x))^2}{2\sigma_y^2(x)} \right] \quad (826)$$
$$= \int dz \left| \frac{dz}{dx} \right|^{-1} P_X(x = g^{-1}(z)) \frac{1}{\sqrt{2\pi\sigma_y^2(z)}} \exp\left[-\frac{(y-z)^2}{2\sigma_y^2(z)} \right], \quad (827)$$

where we have changed variables to $z = g(x)$, which is allowed if the input/output relation is monotonic. But now we can view the integral as an average over a distribution of z, and we know that if the noise is small we can always write

$$\int dz\, F(z) \frac{1}{\sqrt{2\pi\sigma_y^2(z)}} \exp\left[-\frac{(y-z)^2}{2\sigma_y^2(z)} \right] \approx F(z=y) + \frac{1}{2}\sigma_y^2(z=y) \frac{d^2F(z)}{dz^2}\bigg|_{z=y} + \cdots, \quad (828)$$

for any function $F(z)$. Keeping just the leading term, at small noise levels we have

$$P_Y(y) \approx \left[\left| \frac{dz}{dx} \right|^{-1} P_X(x = g^{-1}(z)) \right]_{z=y}. \quad (829)$$

This looks complicated, but it's not. In fact it is the same as ignoring the noise all together and saying that there is some deterministic transformation from x to y, $y = g(x)$, in which case we must have

$$P_X(x)dx = P_Y(y)dy. \quad (830)$$

By the same reasoning, we can also view the variance $\sigma_y^2(x)$ as being a function not of the input x but rather of the output y, so we'll write $\sigma_y^2(y)$.

Problem 157: Details of the small noise approximation, part one. Show that Eq (830) really is the same as Eq (829).

In the small noise approximation, then, the mutual information between x and y thus can be written as

$$I(y;x) \approx -\int dy\, P_Y(y) \log_2 P_Y(y) - \frac{1}{2\ln 2} \int dy\, P_Y(y) \ln[2\pi e \sigma_y^2(y)]. \quad (831)$$

Now it's clear that, given the noise level, we can maximize the mutual information by varying the distribution of outputs $P_Y(y)$. Notice that we started with the problem of varying the distribution of inputs, but now things are formulated in terms of the distribution of outputs; Eq (830) tells us that these are equivalent in the low noise limit. To do the optimization correctly, however, we have to add a Lagrange multiplier that fixes the normalization of the distribution. Thus we are interested in the functional

$$\tilde{I} \equiv I(y;x) - \mu \int dy\, P_Y(y). \quad (832)$$

As usual, to optimize we set the derivative equal to zero:

$$\left.\frac{\delta \tilde{I}}{\delta P_Y(y)}\right|_{P_Y(y)=P_{\rm opt}(y)} = 0 \tag{833}$$

$$\Rightarrow 0 = -\frac{1}{\ln 2}\left[\ln P_{\rm opt}(y) + 1\right] - \frac{1}{2\ln 2}\ln[2\pi e\sigma_y^2(y)] - \mu \tag{834}$$

$$\ln P_{\rm opt}(y) = -\frac{1}{2}\ln[2\pi e\sigma_y^2(y)] - (1 + \mu \ln 2) \tag{835}$$

$$P_{\rm opt}(y) = \frac{1}{\sqrt{2\pi e\sigma_y^2(y)}} e^{-(1+\mu\ln 2)}. \tag{836}$$

We can write this more simply by gathering together the various constants,

$$P_{\rm opt}(y) = \frac{1}{Z}\frac{1}{\sigma_y}, \tag{837}$$

where Z must be chosen so that the distribution is normalized, so

$$Z = \int \frac{dy}{\sigma_y}. \tag{838}$$

With this result for the optimal distribution, the mutual information is

$$I_{\rm opt} = \log_2\left[\frac{Z}{\sqrt{2\pi e}}\right]. \tag{839}$$

Problem 158: Extrema of the mutual information. Once again we need to check that we have found an optimum, rather than some other type of extremum, in the dependence of the mutual information on the distribution of outputs, Eq (831). You can do this explicitly by computing second (functional) derivatives, or by appealing to general convexity properties of the entropy. Notice that our ability to write the information so simply as a functional of the output distribution alone is a feature of the low noise approximation. More generally, we should view the mutual information as a functional of the input distribution $P(x)$ and the conditional distribution(s) $P(y|x)$. Show that, in this more general setting, once $P(y|x)$ is known, the mutual information has a well defined maximum as a functional of $P(x)$.

Problem 159: Details of the small noise approximation, part two. Carry out the small noise approximation to the next leading order in the noise level σ_y^2. Step by step, you should find $P(y)$ and then an expression for the information $I(y;x)$. What can you say about the problem of optimizing $I(y;x)$ in this case?

The result for the optimal distribution of outputs, Eq (837), is telling us something sensible: we should use the different outputs y in inverse proportion to how noisy they are. Suppose, however, that the noise level is constant. Then what we find is that the distribution of outputs should be uniform. How can the system do this? Recall that in the low noise limit, the relationship between input and output is nearly deterministic, so we have Eq (830), $P_Y(y)dy = P_X(x)dx$. But we also have that $y = g(x)$, in this approximation. If $P_Y(y)$ is uniform, this means that

$$P_Y(y) = \frac{1}{y_{\rm max} - y_{\rm min}}, \tag{840}$$

and hence

$$\frac{dy}{dx} = \frac{dg(x)}{dx} = (y_{\rm max} - y_{\rm min})P(x) \tag{841}$$

$$g(x) = (y_{\rm max} - y_{\rm min})\int_{x_{\rm min}}^{x} dx'\, P(x'). \tag{842}$$

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

Problem 160: How general is Eq (842)? We have derived Eq (842) by assuming that the noise is additive, Gaussian, small, and finally has a variance that is constant across the range of inputs or outputs. Show that you can relax the assumption of Gaussianity (while keeping the noise small, additive, and independent of inputs) and still obtain the same result for the optimal input/output relation.

Equation (842) makes clear that any theory which involves optimizing information transmission or efficiency of representation inevitably predicts that the input/output relation must be matched to the statistics of the inputs. 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. Although this is obviously an over–simplified problem, it is tempting to test the predictions, and this is exactly what Laughlin did in the context of the fly's visual system.

Laughlin built an electronic photodetector with aperture and spectral sensitivity matched to those of the fly retina, and used this to scan natural scenes, sampling the distribution of input light intensities $P(\mathcal{I})$ as it would

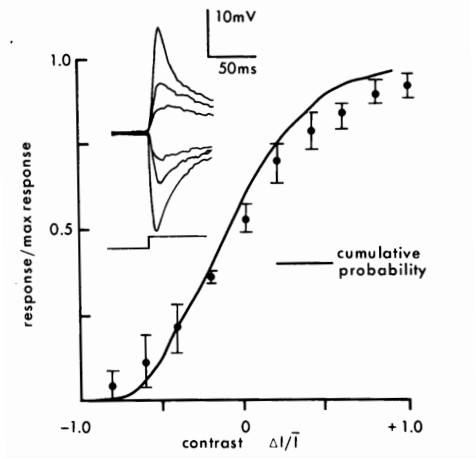

FIG. 149 Input/output relations of large monopolar cells compared with the prediction of Eq (842), from Laughlin (1981). Brief changes in light intensity relative to a mean background produce transient voltage changes in the LMCs (inset), and the peaks of these responses are taken as the cell's output. Normalized responses are compared to the cumulative probability distribution of light intensities, as described in the text.

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, and which we have seen before in [pointers!]—by measuring the peak voltage response to flashes of light. The agreement with Eq (842) was remarkable, as shown in Fig 149, especially when we remember that there are no free parameters. While there are obvious open questions, this is a really beautiful result that inspires us to take these ideas more seriously.

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 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 [show a figure that illustrates this!].

As I have emphasized before, the problems of signals, noise and information flow in the nervous system have analogs within the biochemical and genetic machinery of single cells. For the simple problem of one input and one output, we can move beyond analogy and actually use the same equations to describe these very different biological systems.

Suppose that we have a single transcription factor that controls the expression of one target gene. Now we can think of the input x as the concentration of the transcription factor, and the output y as the expression level of the gene. As in Laughlin's discussion of the fly retina, we are (perhaps dangerously) ignoring dynamics. In the context of gene regulation this probably is best seen as a quasi–steady state approximation, in which the changes in transcription factor concentration are either slow or infrequent, so that the resulting gene expression level has a chance to find its appropriate steady level in response.

We have discussed the problems of noise in the control of gene expression in Section II.B, and a crucial feature of that discussion is that the noise levels cannot be constant. In the simplest case, we are counting molecules, and counting zero molecules allows for no variance, while counting the maximum number of molecules leaves lots of room for variation. For the problem at hand, this means—because of Eq (837)—that the distribution of outputs that maximizes information transmission can't be uniform. To find the form of the optimal distribution we need to recall some of our earlier discussion about noise.

We have identified (at least) three noise sources in the regulation of gene expression. One term is the shot noise in the synthesis and degradation of the mRNA or protein (output noise). The second is the randomness in the arrival of transcription factor molecules at their target site (input noise), and the third is from the kinetics of the

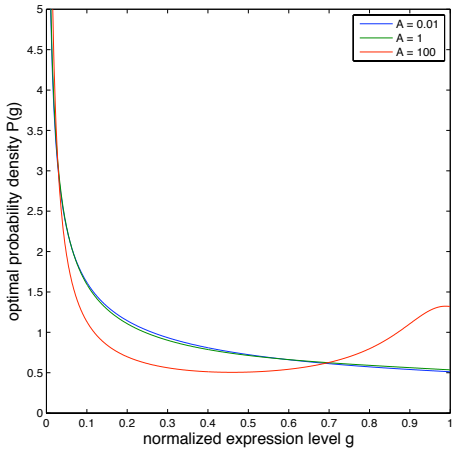

FIG. 150 Optimal distributions of (output) gene expression levels. As described in the text, we maximize the transmission of information from a single transcription factor to a single target gene. Different curves correspond to relative contributions from input and output noise, as in Eq (845).

'switching' events that occur on binding of the transcription factors. We have argued that cells can reduce the impact of this last term by proper choice of parameters, leaving two fundamental sources of noise. The shot noise generates a variance at the output proportional to the mean, while the random arrivals are equivalent to a fluctuation in input concentration $(\delta c/c)^2 \propto 1/c$. Putting these together we have [from the discussion leading to Eq (376)] the variance in the expression level

$$\sigma_g^2(c) = \alpha \bar{g}(c) + \frac{B}{c} \cdot \left| \frac{d\bar{g}(c)}{d \ln c} \right|^2, \tag{843}$$

where α and B are constants, and $\bar{g}(c)$ is the mean expression level as a function of the input transcription factor concentration c; as usual we will normalize the measurements of expression levels so that the maximum $\bar{g}(c) = 1$. Finally, if we can assume that the input/output relation is well approximated by a Hill function,

$$\bar{g}(c) = \frac{c^n}{c^n + K^n}, \tag{844}$$

then we can write the variance as a function of the mean, as in Eq (376),

$$\sigma_g^2(\bar{g}) = \alpha \bar{g} + \beta \bar{g}^{2-1/n} (1 - \bar{g})^{2+1/n}. \tag{845}$$

The parameter $A = \beta/\alpha$ measure the relative importance of input and output noise; large A means that the input noise is dominant near the midpoint of the input/output relation.

In Figure 150 we see the results for the optimal distributions of expression levels, derived using the general result of Eq (837) with the noise variance from Eq (845). We hold the cooperatvity fixed ($n = 5$) and consider what happens as we change the relative importance of the input and output noise (A). As long as output noise is dominant, $A < 1$, the optimal distribution is monotonically decreasing. If we take the results seriously, the distribution has a singularity as we approach zero expression level. There is no physical reason why this can't happen, but we also can't trust our calculation here since at some point the noise $\sigma \propto \sqrt{\bar{g}}$ will become larger than the mean as $\bar{g} \to 0$. Nonetheless, it's clear that when output noise is dominant, the optimal distribution of expression levels is relatively featureless, biased toward low expression levels. The strength of this bias is considerable so that the probability of having more than half–maximal activation, $\int_{1/2}^1 dg\, P(g)$, is a bit less than 30%.

At larger values of A, where input noise is more important, the optimal distribution of expression levels becomes bimodal. This is especially interesting, because extreme bimodality corresponds to a simple on/off switch. Intuitively, true switch–like behavior runs counter to the idea that information transmission is being maximized:

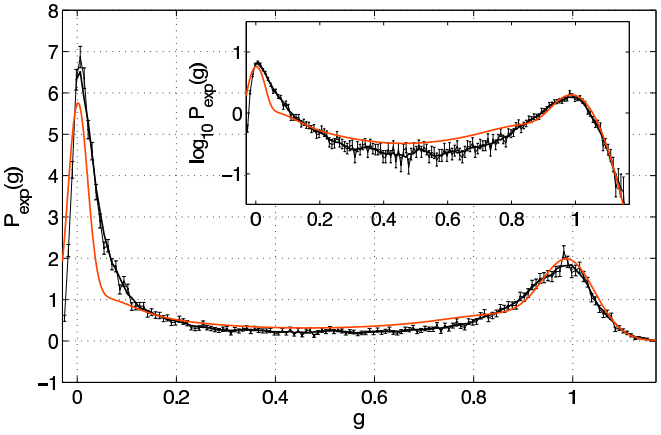

FIG. 151 Distributions of Hunchback expression levels in the early fruit fly embryo (Tkačik et al 2008). In red, the distribution predicted by optimizing information transmission given the measured input/output relation and noise in the control of Hb by Bcd. In black, with error bars, the distribution measured experimentally.

we might expect that maximizing information transmission involves making extensive use of intermediate expression levels, while building a reliable switch means exactly the opposite, avoiding intermediate levels. In fact, few of classic examples of "genetic switches" are perfect, and here we see that maximizing information transmission can lead to relatively low probabilities of occupying intermediate levels, just depending on the structure of the noise in the system.

We can bring this theoretical discussion down to earth by considering a real system. As discussed in Section II.B, there are measurements on the input/output relation and noise level for the control of the *hunchback* gene by the transcription factor Bicoid in the early *Drosophila* embryo. If we take the formalism above seriously, we can use these measurements to predict, with no free parameters, the distribution of *hunchback* expression levels, which can also be extracted from the experiments. To do this correctly, we should go beyond the small noise approximation and solve the full optimization problem numerically; the results are shown in Fig 151.

Figure 151 is the direct analog of Laughlin's result in the fly retina. As in that case, the agreement of theory and experiment is very good, and again it should be emphasized that there are no free parameters—these are not models we are fitting to data, but quantitative predictions from theory. One can go further, and show from the data that the actual amount of information[86] being transmitted from Bicoid to Hunchabck is 0.88 ± 0.09 of

[86] This is a good place to remember the technical difficulties involved in estimating information from finite samples of data. See Appendix A.9.

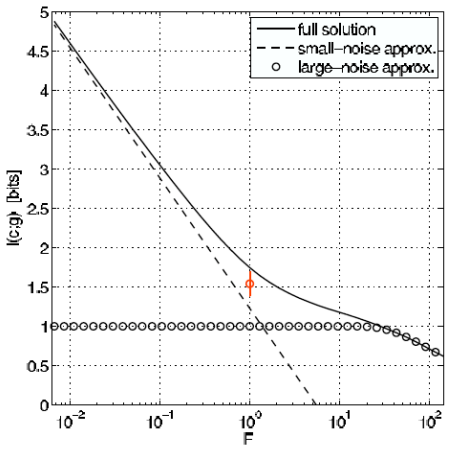

FIG. 152 Changes in information transmission from Bcd to Hb if we scale all noise variances by a factor F, from Tkačik et al (2008). This is equivalent to scaling all the numbers of molecules by a factor $1/F$. At each noise level we compute the maximum information transmission, as described in the text. Limiting behaviors in the small and large noise approximations are shown for reference. The real system (in red, with error bar) is in an intermediate regime, although close to the small noise limit.

the limit set by the measured noise levels. Thus, going back to the remarks in Section 1.5, we can see directly that the system is operating near its optimum. This optimum corresponds to significantly more than one bit, which means that intermediate expression levels, beyond an on/off switch, are being used reliably. Finally, since we understand how the absolute numbers of molecules influence the noise level in the system (see, again, Section II.B), we can compute that more bits would be very expensive—doubling the information would require twenty times as many molecules, as shown in Fig 152.

Problem 161: Information flow through calcium binding proteins. Many biological processes are regulated by calcium. Typically the regulatory process begins with calcium binding to a protein. In almost all cases, there are multiple binding sites, and these sites interact cooperatively. We'd like to understand something about the signals, noise and information flow in such regulatory systems; not much has been done in this aream so this is a deliberately open ended problem. Consider the simple model shown in Fig 153. This is a dimeric protein with four states, corresponding to empty and filled Ca^{++} binding sites on each of the two monomers. The sites interact, since the rate of unbinding from one site depends on the occupancy of the other site.

(a.) Calculate the equilibrium probability of occupying each of the states in Fig 153. Use these results to plot the fraction of occupied binding sites as a function of the calcium concentration c. You should be able to choose units which eliminate all parameters except for the dimensionless constant F. Show that for $F = 1$ your results are equivalent to having two independent binding sites, and that the fraction of occupied sites becomes more strongly sigmoidal or switch–like as F becomes larger. Cooperativity means, in this context, that the free energy change upon binding of a calcium ion to one site is increased by occupancy of the other site. Relate the parameter F to this free energy difference or interaction energy. Can interaction energies of just a few times k_BT make a difference in the shape of the plot of occupancy vs. concentration? See also the discussion of cooperativity in Appendix A.4.

(b.) Suppose that we have N copies of this protein in the cell, all experiencing the same calcium concentration. Let the number of molecules with no bound calcium be n_0, the number with one bound calcium be n_1, and the number with two bound calcium be n_2; of course $\sum_j n_j = N$. Use your results from Problem 1 to calculate the mean values of each n_j and the covariance matrix $C_{jk} = \langle \delta n_j \delta n_k \rangle$. Verify that the determinant of the covariance matrix is zero in this formulation. Why is this true? Notice that we're only asking here about the fluctuations that you would see in a single snapshot of the molecules, not about the dynamics or spectrum of this noise.

(c.) It is widely assumed that in systems such as this, only the state with full occupancy of the binding sites is really "active." In practice what this means is that the calcium binding protein is associated with some other protein, such as an enzyme, and the enzyme becomes active only when both Ca^{++} are bound. Thus, the output of the system is something proportional to n_2. Calculate the change in the mean $\langle n_2 \rangle$ that results from a small change in calcium concentration $c \to c + \delta c$. Compare this with the variance $\langle (\delta n_2)^2 \rangle$ to compute a signal–to–noise ratio, or the equivalent noise level $\delta c_{\rm rms}$ in the calcium concentration itself. Plot your results. Again, you should be able to put everything into unitless form, leaving only the parameter F. Does making the system more switch–like by increasing F makes it more sensitive to small changes in concentration, as you might expect? Are there competing effects which could result in better performance at smaller F?

(d.) Suppose that molecules with one bound calcium also are active. Then the output activity of the system is proportional to some mixture of n_1 and n_2, which we can write as $A = (1-a)n_2 + an_1$; note that $a = 0$ brings us back to the case where only doubly–bound states are active. Compute sensitivity of the mean activity, $\partial\langle A \rangle/\partial c$ and the variance $\langle (\delta A)^2 \rangle$. If the system is operating at a particular calcium concentration c, can you lower the effective noise level

$$\delta c_{\rm rms} \equiv \sqrt{\langle (\delta A)^2 \rangle} \left| \frac{\partial \langle A \rangle}{\partial c} \right|^{-1}$$

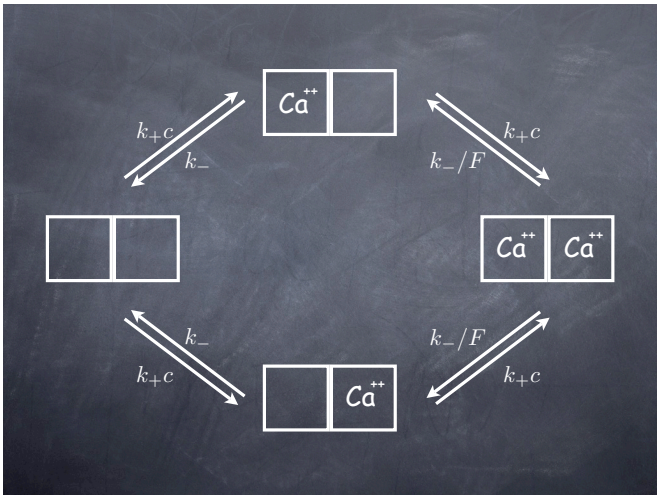

FIG. 153 Model of calcium binding to a dimeric protein. The rate at which calcium binds to each site, k_+ is assumed to be the same and independent of the occupancy of the other site. The unbinding rates, however, are different depending on whether the other site is empty (k_-) or filled (k_-/F).

by choosing $a \neq 0$? Can you lower the noise level at all calcium concentrations using the same value of a, or are there tradeoffs?

(e.) One way to think about the effective noise level $\delta c_{\rm rms}$ is that it sets a scale for the smallest concentration differences that can be detected. If we imagine that c can range from zero up to some maximum $c_{\rm max}$, then it seems natural to say that number of different levels of concentration that can be distinguished is given by

$$N_{\rm levels} = \int_0^{c_{\rm max}} \frac{dc}{\delta c_{\rm rms}(c)}, \tag{846}$$

where we note explicitly that the noise level depends on the background concentration. The number of distinguishable levels should translate into an information transmission (in bits) of $I \sim \log_2(N_{\rm levels})$, and this is almost right in the limit that the noise is small. Show how a rigorous version of this argument can be constructed by analogy with the derivation of Eq's (838) and (839). Calculate I for the system discussed above. Does thinking about the information transmitted, rather than just the noise level, help you to decide whether there is a uniquely best mixture of activity from the singly– and doubly–bound states? What is the impact of the cooperativity (here captured by the parameter F) on the information transmission?

(f.) Some things to explore: Your results above suggest that, at least under some conditions, it would be useful if the system "reads out" some combination of the singly– and doubly–occupied states. Can you find hints in the literature of the predicted partial activation? For concreteness, focus on the case of calmodulin. Our discussion above is for snapshots of the molecules, so 'noise' just means the total variance. Suppose that the readout scheme effectively averages over a time longer than the times required for transitions among the different states. Then you need to compute the spectral density of the noise, and follow the path we discussed in the context of bacterial chemotaxis. Is there anything qualitatively new here, or just a change in details?

One of many questions left open in Laughlin's original discussion is the time scale on which the matching should occur. One could imagine that there is a well defined distribution of input signals, stable on very long time scales, in which case the matching could occur through evolution. Another possibility is that the distribution is learned during the lifetime of the individual organism, perhaps largely during the development of the brain to adulthood. Finally one could think about mechanisms of adaptation that would allow neurons to adjust their input/output relations in real time, tracking changes in the input distribution. It seems likely that the correct answer is all of the above. But the last possibility, real time tracking of the input distribution, is interesting because it opens the possibility for new experimental tests.

We know that some level of real time matching occurs, as in the example of light and dark adaptation in the visual system. We can think of this as neurons adjusting their input/output relations to match the mean of the input distribution. The real question, then, is whether there is adaptation to the distribution, or just to the mean. Actually, there is also a question abut the world we live in, which is whether there are other features of the distribution that change slowly enough to be worth tracking in this sense.

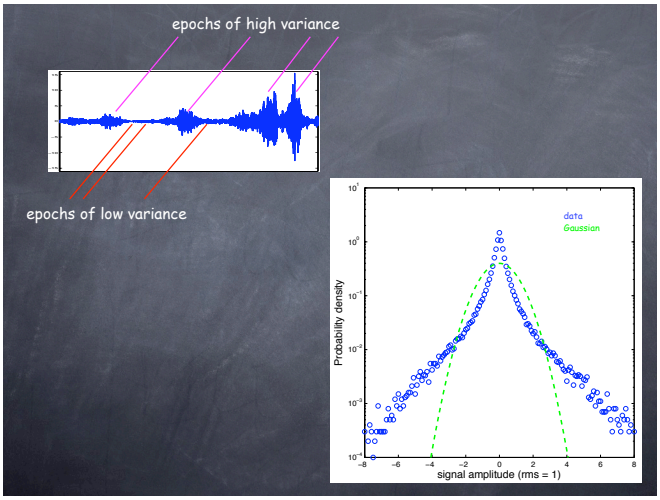

FIG. 154 This is a placeholder .. should replace with real data, e.g. sounds from the songbird colony. Intermittency in natural sounds. Top trace shows the alternating "loud" and "soft" period characteristic of natural sounds. Probability distribution of the instantaneous signal amplitude is far from Gaussian, having long, nearly exponential tails. Need to illustrate more clearly that these tails are removed by local variance normalization!

As an example, we know that many signals that reach our sensory systems come from distributions that have long tails (cf Fig 154). In some cases (e.g., in olfaction, where the signal—odorant concentration—is a passive tracer of a turbulent flow) there are clear physical reasons for these tails, and indeed its been an important theoretical physics problem to understand this behavior quantitatively. In most cases, the tails arise through some form of intermittency. Thus, we can think of the distribution of signals as being approximately Gaussian, but the variance of this Gaussian itself fluctuates; samples from the tail of the distribution arise in places where the variance is large. This scenario also holds for images of the natural world, so that there are regions of high variance and regions of low variance. The possibility of such "variance normalization" in images suggests that the visual system could code more efficiently by adapting to the local variance, in addition to the local mean (light and dark adaptation).

Adaptation to local variance, or more generally adaptation to input statistics beyond the mean, definitely happens at many stages of neural processing (Fig 155). The earliest experiments looked the responses of retinal ganglion cells to sudden changes in the variance of their inputs, and showed that there is a pattern very similar to what one sees with sudden changes in mean. More ambitious experiments on the motion–sensitive neurons in the fly visual system mapped the input/output relation when inputs were drawn from different distributions, and found that the input/output relation scales in proportion to the dynamic range of inputs, which is what

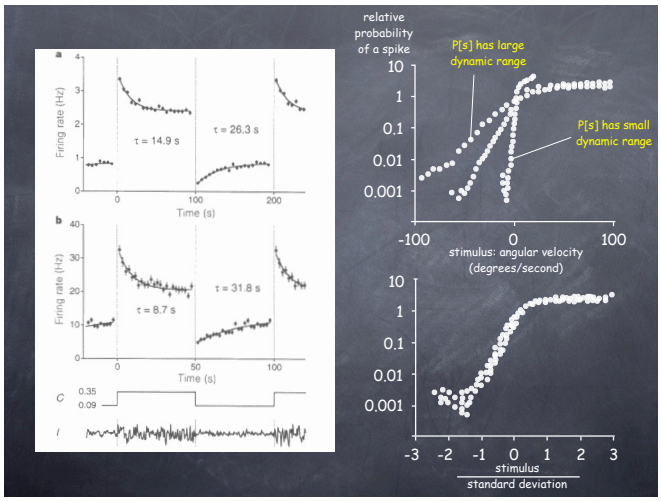

FIG. 155 At left, adaptation of retinal ganglion cells to sudden changes in the variance of light intensity (Smirnakis et al 1997). At right top, input/output relations for the fly motion–sensitive neuron H1 measured when inputs are drawn from different distributions (Brenner et al 2000). To be precise one has to define the input as a filtered version of the velocity, and the methods for determining these filters are discussed in Appendix A.7. At right bottom, the input/output relations collapse when expressed a function of the stimulus in units of its standard deviation.

one expects from the matching principle if noise levels are small; it was also checked that the precise proportionality constant in the scaling relation served to maximize information transmission. Further, if you suddenly switch from one distribution to another, you can 'catch' the system using the wrong code and transmitting less information, but the adaptation to the new distribution is very fast, close to the limit set by the need to collect enough samples that you are sure there was a change. Related observations have been in many systems, from low level sensory neurons up to mammalian cortex. [Do we want to say more here? Maybe work on how such adaptation is a property of individual neurons, so it is a building block of neural computation? At least pointers to the fact that these effects are so fast that calling them "adaptation" raises some questions.]

I think the adaptation experiments are important because they give a whole new way of testing the ideas about matching between the input/output relation and the distribution of inputs—by changing the input distribution, if you believe the theory, we should drive changes in the input/output relation, and it seems that this works. Can we imagine a similar experiment in the genetic or biochemical systems? In truth, there are few cases (aside from embryonic development) where we have quantitative measurements on the distributions of inputs under moderately natural conditions. If we change the distribution, then for the case of gene regulation one imagines that input/output relations could change in response only on evolutionary time scales, but at least for bacteria such evolutionary experiments are now quite feasible. Certainly there are models for network evolution that use information theoretic quantities as a surrogate for fitness, and these models are generating interesting predictions, as shown in Fig [include a figure from Francois & Siggia simulations]. It would be exciting to see laboratory evolution experiments that are the analog of the neural experiments in Fig 155.

So far the discussion is about one input and one output, in single genes or neurons. Almost all the really interesting systems, however, involve populations or networks of these elements. Indeed, one of the earliest ideas about optimizing information transmission in neural coding is that interactions among neighboring neurons in the retina serve to reduce the redundancy of the signals that they transmit, thus making better use of their capacity.

To get a feeling for how redundancy reduction works, consider a system in which there are N receptor cells that produce signals x_i, and these feed into a layer of N output neurons that take linear combinations of their inputs and add noise, so that the outputs of the system are

$$y_i = \sum_j W_{ij} x_j + \eta_i, \tag{847}$$

and shown schematically in Fig 847. In the simplest case the noise will be Gaussian and independent in each output neuron, $\langle \eta_i \eta_j \rangle = \delta_{ij}\sigma^2$. Let's also assume, again for simplicity, that the distribution of the xs is also Gaussian, with zero mean and a covariance matrix $\langle x_i x_j \rangle = C_{ij}$. Then following the arguments in Section IV.A, the information that the outputs provide about the inputs is

$$I(\vec{y};\vec{x}) = \frac{1}{2}\mathrm{Tr}\log_2\left(\mathbf{1} + \frac{1}{\sigma^2}WCW^T\right), \tag{848}$$

where W^T denotes the transpose of the matrix W. We can chose the matrix W, which defines the "receptive fields" of the output neurons [point back to first discussion of receptive fields; check!] to maximize the information, but we need a constraint, since other wise the answer is always to make W larger so we can overwhelm the noise. A natural constraint, then, is to fix the overall dynamic range of the output signal,

$$\sum_i \langle y_i^2 \rangle = \mathrm{Tr}\left(WCW^T\right) + N\sigma^2. \tag{849}$$

But if we go into a basis where WCW^T is diagonal, then the information becomes

$$I(\vec{y};\vec{x}) = \frac{1}{2}\sum_\mu \log_2\left(1 + \frac{\Lambda_\mu}{\sigma^2}\right), \tag{850}$$

where the Λ_μ are the eigenvalues of the matrix WCW^T, and the constraint is that the sum of these eigenvalues

must be constant. Then it is clear from the convexity of the logarithm that the best we can do is to have all of the eigenvalues be equal, which means that WCW^T is proportional to the unit matrix. But (leaving aside the contribution from the noise), WCW^T is the correlation matrix of the output signals. Thus, in this simple model, we maximize information transmission by removing all of the correlations in the input, and making the outputs independent of one another.

Problem 162: Convexity and equalization. Show explicitly that if we want to maximize

$$I = \frac{1}{2}\sum_\mu \log_2(1+\tilde{\Lambda}_\mu), \tag{851}$$

subject to the constraint

$$\sum_\mu \tilde{\Lambda}_\mu = C, \tag{852}$$

then the solution is to have all the $\{\tilde{\Lambda}_\mu\}$ be equal, that is $\tilde{\Lambda}_\mu = \Lambda_0$. [I'd like to get the students to think more about the implications of this ...]

In the retina, we expect that correlations, and perhaps also the transformations from input to output, are translation invariant. Thus if the receptor cell i is at position $\mathbf{r}_\mathrm{i}$, perhaps on a lattice, and the output neurons are on the same lattice, we should have

$$C_\mathrm{ij} = C(\mathbf{r}_\mathrm{i} - \mathbf{r}_\mathrm{j}), \tag{853}$$
$$W_\mathrm{ij} = W(\mathbf{r}_\mathrm{i} - \mathbf{r}_\mathrm{j}). \tag{854}$$

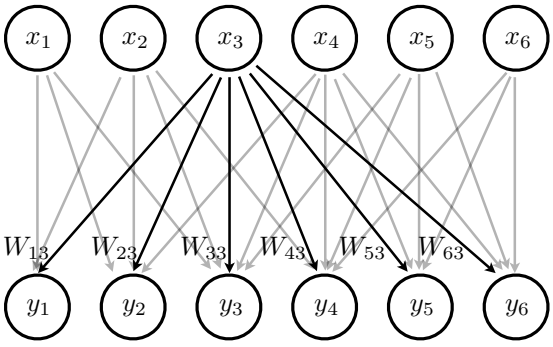

FIG. 156 A schematic network, after Eq (847). The x_j provide inputs to the y_i, with weights W_ij. All connections are present, but the connections from x_3 are highlighted.

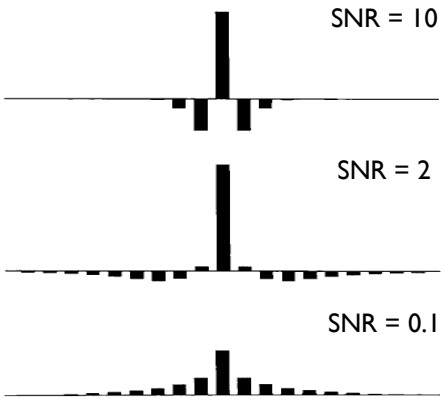

FIG. 157 Cross–sections through the optimal matrices W_ij in the problem with noise, from Atick & Redlich (1990). The correlation function is assumed to be exponential, $C_\mathrm{ij} \propto \exp(-|\mathrm{i}-\mathrm{j}|/\xi)$, with $\xi = 50$, much longer than the range of interactions shown here. At high SNR, the solution looks like a differentiator, which decorrelates the signals, while at low SNR the solution integrates to suppress noise.

Then the condition for independence at the outputs becomes

$$\delta_\mathrm{ij} \propto \sum_\mathrm{km} W_\mathrm{ik} C_\mathrm{km} W_\mathrm{jm} \tag{855}$$
$$= \sum_\mathrm{km} W(\mathbf{r}_\mathrm{i} - \mathbf{r}_\mathrm{k}) C(\mathbf{r}_\mathrm{k} - \mathbf{r}_\mathrm{m}) W(\mathbf{r}_\mathrm{j} - \mathbf{r}_\mathrm{m}). \tag{856}$$

We approximate the sums as integrals, so that

$$\delta_\mathrm{ij} \approx \int d^2r' \int d^2r''\, W(\mathbf{r}_\mathrm{i} - \mathbf{r}') C(\mathbf{r}' - \mathbf{r}'') W(\mathbf{r}_\mathrm{j} - \mathbf{r}'') \tag{857}$$
$$= \int \frac{d^2k}{(2\pi)^2} |\tilde{W}(\mathbf{k})|^2 S(\mathbf{k}) e^{i\mathbf{k}\cdot(\mathbf{r}_\mathrm{i} - \mathbf{r}_\mathrm{j})}, \tag{858}$$

where $\tilde{W}(\mathbf{k})$ is the Fourier transform of $W(\mathbf{r})$, and we identify the Fourier transform of the correlation function $C(\mathbf{r})$ as the power spectrum $S(\mathbf{k})$. To satisfy this condition $|\tilde{W}(\mathbf{k})|^2 S(\mathbf{k})$ must be constant, independent of $\mathbf{k}$, and if $W(\mathbf{r})$ is symmetric in space this means that

$$\tilde{W}(\mathbf{k}) \propto \frac{1}{\sqrt{S(\mathbf{k})}}. \tag{859}$$

We expect that the power spectrum of correlations in the inputs to the retina fall off at high frequencies, which means that the optimal weights W have the form of a filter which does the opposite, attenuating the low frequencies and enhancing high frequencies. In fact, experiments show that the power spectrum of contrast in natural scenes is scale invariant, so that $S(\mathbf{k}) \propto |\mathbf{k}|^{-\alpha}$, with the exponent α close to 2. Then the optimal weights W

should actually vanish as $\mathbf{k} \to 0$, which means that the output of the retina should be insensitive to spatially uniform illumination; on the other hand, the output should overemphasize gradients or edges. By including the effects of noise (as in the problem below), one can see a crossover to spatial averaging at low SNR; see Fig 157. Qualitatively this is all correct: at high signal–to–noise ratios we can see this enhancement of edges not just in the responses of retinal ganglion cells but also in our perception, through the phenomenon of Mach bands [figures?], and this spatial differentiation gives way to integration as we lower the light levels and hence the SNR.

Problem 163: Redundancy reduction vs noise reduction. Equation (859) suggests that at large $\mathbf{k}$, where the power spectrum of input signals should be small, the weight in transferring these signals to the output should be large. This can't be completely right, since we expect that at very high (spatial) frequencies, signals will be lost in a background of noise. Go back to the start of this analysis and assume that the signals x_i already have a little bit of noise attached to them (as with photon shot noise in vision) so that

$$y_i = \sum_j W_{ij}(x_j + \xi_j) + \eta_i, \tag{860}$$

where everything is as before but $\langle \xi_i \xi_j \rangle = \delta_{ij}\sigma_0^2$. Follow the outline above and derive the form of the weights W_{ij} that optimize information transmission at fixed output variance. Verify that as $\sigma_0 \to 0$ you recover the simple picture in which the optimal W_{ij} serve to remove correlations. Show, in contrast, that as σ_0 becomes large, the optimal solution involves averaging over multiple inputs to beat down the noise.

Problem 164: Information available at the retina. Give a problem that takes the students through the calculation in Ruderman & Bialek (1994), showing that with reasonable assumptions natural scenes provide only ~ 1 bit per cone in the fovea.

[Do we want to talk about coding/whitening in the time domain, maybe the results on filtering at the receptor/LMC synapse? Could argue by analogy with spatial whitening, give a problem to work out details.]

Maybe a simpler example of these ideas is provided by color processing. Roughly speaking, at one point in space our retina takes three samples, corresponding to the signals in the three different cones. These three signals are correlated, both because the absorption spectra of the pigments in the different cones overlap and because the reflectance spectra of the objects around us are rather smooth functions of wavelength.[87] By analogy with what we have seen thus far, if the retina is under pressure to maximize information transmission then it should send these signals to the brain in some decorrelated form. Early guesses about the form of the correlations among the different cone signals suggested that the three decorrelated signals would correspond roughly to the sum of all the inputs (the total light intensity, ignoring color), an approximately "red minus green" signal, and a "blue minus yellow" signal. In fact it is known that neurons throughout the visual system follow this pattern of "opponent" color processing [feels like there should be something about experiments demonstrating color opponency, but I don't know how quantitatively one can make comparisons, so ...??].

To do a more quantitative analysis one has to get away from traditional color photography, because (for example) the three channels in a CCD camera don't have wavelength sensitivities that correspond exactly to that of our cones. Instead one can take hyperspectral images, essentially measuring the spectrum of light at each point in the scene, and then construct the expected signals that will be seen by each cone, known the absorption spectra of the three cone pigments. This analysis shows, quite remarkably, that the rough intuition about opponent processing is nearly exact, with the decorrelated signals being almost perfect integer combinations of the cone signals: if the three cone signals are $\mathcal{L}$, $\mathcal{M}$ and $\mathcal{S}$ for the long, medium and short wavelengths, then the decorrelated signals are $\ell = \mathcal{L} + \mathcal{M} + \mathcal{S}$ (light intensity), $\alpha = \mathcal{L} + \mathcal{M} - 2\mathcal{S}$ (blue minus yellow), and $\beta = \mathcal{L} - \mathcal{M}$ (red minus green), where the coefficients are unity with an accuracy of $\sim 1\%$. Further, this linear transformation serves to generate truly independent signals, even though the underlying distributions are not Gaussian; see Fig 158. These very clean results come from a delicate interplay between the statistical structure of the world and the properties of our visual pigments. I don't know how accurately the coefficients in opponent color processing have been measured, but this is a striking prediction that certainly captures the qualitative behavior of the system and deserves to be tested more quantitatively.

In the example of Fig 157, the optimal weights for transforming receptor signals into neural output correspond to a "center–surround" structure in which an output neuron at point $\mathbf{r}$ gives a positive weight to the receptor cell at point $\mathbf{r}$, and a negative weight to its neighbors. Alternatively, we can think that all weights from receptors to neurons are positive, and the output neurons inhibit one another before sending their signals on to the brain. In our retinae things are complicated, because the transformation from photoreceptors to ganglion cells involves several intermediate cells, but in some simpler creatures such as the horseshoe crab the picture of "lateral inhibition" seems to be correct, and indeed the horseshoe crab was the first retina in which receptive fields were measured. Lateral inhibition is thought to be a general neural mechanism for "sharpening" the responses to stimuli that vary across an array of neu-

[87] In fact this is the same effect. The reflectance properties of most naturally occurring objects in our terrestrial environment are determined by the absorption spectra of organic pigments, and these tend to be broad; see Section [**] and Appendix [**].

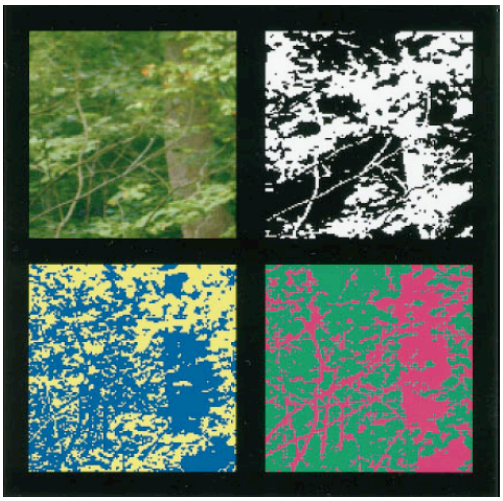

FIG. 158 Statistical structure of color images, from Ruderman et al (1998). Top left shows a color image of one scene analyzed by hyperspectral imaging. From the raw data, one constructs the signals $\mathcal{L}$, $\mathcal{M}$ and $\mathcal{S}$ corresponding to the (log) photon capture rates by each of the types of cones, and then rotates into the basis defined by ℓ (upper right), α (lower left), and β (lower right), as described in the text; images are shown in these three projections after thresholding for clarity. The three images are uncorrelated.

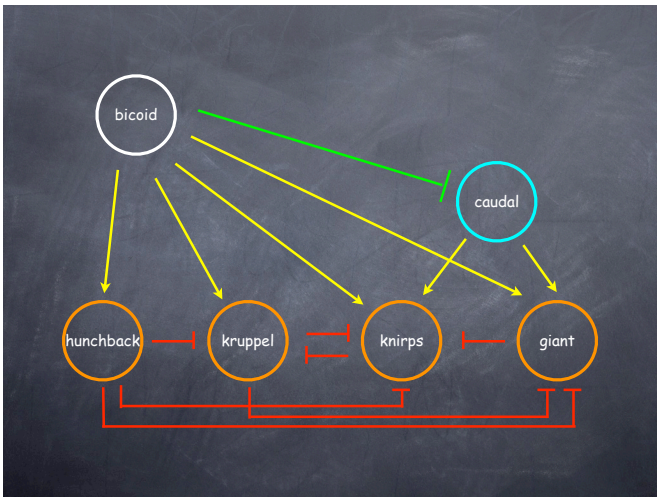

FIG. 159 The gap gene network in the *Drosophila* embryo. Need to check and see how much of this has been discussed already, although this is a nice place to put this ... .

rons, and we have seen that this sharpening is essential in decorrelating signals and enhancing the efficiency of information transmission. Could there be an analog of this for the transmission of information through genetic or biochemical networks? If we go back to the case of Bicoid regulating the expression of Hunchback, we know that this is just one piece of a larger network in which the primary morphogen Bicoid feeds into a collection of gap genes, which in turn interact with one another. Because transcription factors tend to be either activators or repressors, in the absence of any other effects all of the target genes would have correlated expression levels and hence provide redundant data about the concentration of the input. This redundancy can be removed by lateral inhibition, and that is what we see in the gap gene network (Fig 159). The challenge is to take this quantitative analogy and turn it into a quantitative theory.

The representations of data constructed by the nervous system might be efficient in the sense we have considered here, but they have a more obvious feature—they are built from discrete action potentials or spikes. If we look with some reasonably fine time resolution, $\Delta\tau < 10\,\text{ms}$, then since the average spike rates are less than $100\,\text{spikes/s}$, at any moment the typical neuron is silent. In this sense, the code is "sparse." It is this sparseness which, among other things, makes it possible to decode spike trains using linear filters, as in Eq (729). Spikes are expensive, requiring substantial energy expenditure, and perhaps it is this cost which drives the brain toward the construction of sparse representations.

If we take the idea of linear reconstruction seriously, then if the sensory input is $s(t)$—for example sound pressure as a function of time in the auditory system—we would like to have a family of neurons labeled by μ that spike at times $\{t_i^\mu\}$ such that

$$s_{\text{est}}(t) = \sum_\mu \sum_i f_\mu(t - t_i^\mu) \tag{861}$$

is as close as possible to the true signal. Notice that in this system the input is $s(t)$ and the output is the set of spikes $\{t_i^\mu\}$. If we imagine adjusting the input/output relations, the mapping $s(t) \to \{t_i^\mu\}$ will change, perhaps in complicated ways. But suppose we knew the functions $f_\mu(\tau)$. Then there would be "best times" t_i^μ for each spike so that the match between $s_{\text{est}}(t)$ and $s(t)$ is as close as possible. We could imagine searching through some large space of input/output relations to find one that puts the spikes at these best times, or we could use the times themselves as our description of the input/outpute relation. Conversely, if we knew the spike times, we could adjust the filters $f_\mu(\tau)$, as in our previous discussions. Can we do both problems, subject to a constraint on the total number of spikes? This is hard, but by slightly softening the problem—allowing each term in Eq (861) to have a varying amplitude—it becomes tractable. [Can we say something about whether the softening makes much difference in the end? Need to ask Lewicki for details.]

Figure 160 shows the results of this approach applied to a small population of neurons "trained" to provide an efficient representation of natural sounds. There are several interesting features in these results. First, the filters $f_\mu(\tau)$ are localized in time; although they are "tuned" to particular frequencies, they are more like a wavelet than a Fourier representation, with support over a window of time that scales inversely with the characteristic fre-

quency. The filters also a very asymmetric shape, with a sharp attack and a slower decay. If we look through measurements of the impulse responses of neurons emerging from the mammalian ear, we see exactly these structures, and one can even find cells that overlay the predicted filters almost perfectly.[88] Importantly, these structures are lost if one tries to build representations of very different sound ensembles.

By allowing for different total numbers of spikes, or limiting the time resolution with which the spikes are placed, one can construct codes of different qualities. For these different codes it is relatively easy to put an upper bound on the entropy of the spike trains, and to measure the errors between $s_{\rm est}(t)$ and the true $s(t)$; putting these together one obtains the rate–distortion curve for this family of codes. Applied to ensembles of human speech, the results are comparable to or better than conventional coding schemes. It does indeed seem that nature has found an efficient class of codes, not just in abstract terms.

Is there more we want to say here? What about predictive information? I'd love to say something about this, if we know enough ... .

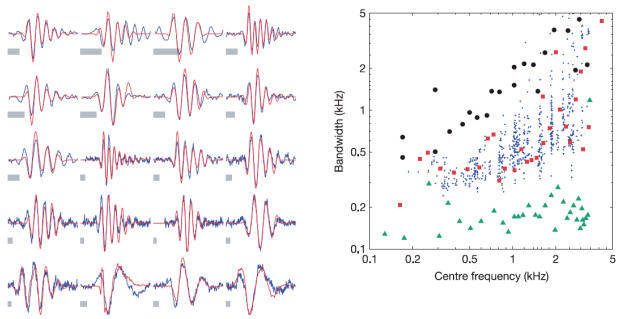

FIG. 160 Ingredients for an efficient representation of natural sounds, as in Eq (861), from Smith & Lewicki (2006). At left, the functions $f_\mu(\tau)$ in red, compared with the impulse responses of single neurons in the cat auditory nerve in blue; grey scale bars are 5 ms long. All these filters are band pass, so that their Fourier transforms $\tilde{f}_\mu(\omega)$ have maximum magnitude at some characteristic frequency and fall to half maximal over some bandwidth. At right, a scatter plot of bandwidths vs characteristic frequencies for the filters (red) and auditory neurons (small blue dots); filters trained on different ensembles (black circles and green triangles) have very different behavior.

[88] One needs to be careful here. The impulse responses of the neurons are measured by the reverse correlation method (see Appendix A.7) which, ideally, extracts a filter characteristic of the *encoding* of sounds into spikes. In contrast, the filtrs $f_\mu(\tau)$ are characteristic of the *decoding* process. One can circumvent this problem by using reverse correlation to analyze the model code, with almost identical results. This suggests that the model, at least, is operating in a regime where the coding and decoding filters are similar. This happens exactly in the limit where all spikes are statistically independent from one another, so that there is no redundancy. Thus, the search for efficient codes may also drive the emergence of simplicity in decoding.

Laughlin's classic paper on matching input/output relations to the distribution of inputs still is very much worth reading, thirty years later (Laughlin 1981). The corresponding analysis for a genetic regulatory element is by Tkačik et al (2008a), with more theoretical exploration in Tkačik et al (2008b). The literature on information transmission in biochemical and genetic networks is growing rapidly; for examples see Ziv et al (2007), Mugler et al (2008), Yu et al (2008) and Tostvein & ten Wolde (2009). For a detailed model of calcium signaling via the protein calmodulin (of relevance to Problem **), see Pepke et al (2010).

Laughlin 1981: A simple coding procedure enhances a neuron's information capacity. SB Laughlin, *Z Naturforsch* **36c,** 910–912 (1981).

Mugler et al 2008: Form, function and information processing in small stochastic biological networks. A Mugler, E Ziv, I Nemenman & CH Wiggins, *IET Sys Biol* (2008).

Pepke et al 2010: A dynamic model of interactions of Ca^{2+}, calmodulin, and catalytic subunits of Ca^{2+}/calmodulin–dependent protein kinase II. S Pepke, T Kinzer–Ursem, S Mihalas & MB Kennedy, *PLoS Comp Biol* **6,** e1000675 (2010).

Tkačik et al 2008a: Information flow and optimization in transcriptional regulation. G Tkačik, CG Callan Jr & W Bialek, *Proc Nat'l Acad Sci (USA)* **105,** 12265–12270 (2008).

Tkačik et al 2008b: Information capacity of genetic regulatory elements. G Tkačik, CG Callan Jr & W Bialek, *Phys Rev E* **78,** 011910 (2008).

Tostvein & ten Wolde 2009: Mutual information between input and output trajectories of biochemical networks. F Tostvein & PR ten Wolde, *Phys Rev Lett* **102,** 21801 (2009).

Yu et al 2008: Negative feedback that improves information transmission in yeast signalling. RC Yu, CG Pesce, A Colman–Lerner, L Lok, D Pincus, E Serra, M Holl, K Benjamin, A Gordon & R Brent, *Nature* **456,** 755–761 (2008).

Ziv et al 2007: Optimal signal processing in small stochastic biochemical networks. E Ziv, I Nemenman & CH Wiggins, *PLoS* **2,** e1007 (2007).

The idea of matching input/output relations to the statistics of inputs had a big impact on neuroscience, in particular driving the exploration of input statistics under natural conditions. An early paper in this direction was by Field (1987), who noted that the power spectra of natural images were approximately scale invariant. Natural images are strongly nonGaussian, so we expect that scaling should mean much more than an appropriately shaped power spectrum, and this is true (Ruderman & Bialek 1994, Ruderman 1994); exploration of these statistical structures beyond the Gaussian approximation led to the ideas of variance normalization, and the prediction of adaptation to the local variance. Well before these analyses there was a substantial body of work on "contrast gain control" at various levels of the visual system; see, for example, Shapley & Victor (1981). The work on image statistics prompted a more explicit search for adaptation to the distribution

of visual inputs beyond the mean light intensity (Smirnakis et al 1997). Brenner et al (2000) describe experiments mapping the input/output relations of the fly's motion–sensitive visual neurons when inputs are drawn from different distributions, demonstrating that this adaptation served to optimize information transmission, and Fairhall et al (2001) explored the dynamics of this process, showing that one could "catch" the system using the wrong code and transmitting less information.

Brenner et al 2000: Adaptive rescaling optimizes information transmission. N Brenner, W Bialek & R de Ruyter van Steveninck, *Neuron* **26,** 695–702 (2000).

Fairhall et al 2001: Efficiency and ambiguity in an adaptive neural code. AL Fairhall, GD Lewen, W Bialek & RR de Ruyter van Steveninck, *Nature* **412,** 787–792 (2001).

Field 1987: Relations between the statistics of natural images and the response properties of cortical cells. DJ Field, *J Opt Soc Am A* **4,** 2379–2394 (1987).

Ruderman 1994: The statistics of natural images. DL Ruderman, *Network* **5,** 517–548 (1994).

Ruderman & Bialek 1994: Statistics of natural images: Scaling in the woods. DL Ruderman & W Bialek, *Phys Rev Lett* **73**, 814–817 (1994).

Shapley & Victor 1981: How the contrast gain control modifies the frequency responses of cat retinal ganglion cells. RM Shapley & JD Victor, *J Physiol (Lond)* **318,** 161–179 (1981).

Smirnakis et al 1997: Adaptation of retinal processing to image contrast and spatial scale. S Smirnakis, MJ Berry II, DK Warland, W Bialek & M Meister, *Nature* **386,** 69–73 (1997).

Adaptation to the distribution of inputs has now been reported in many neural systems: in the song bird (Nagel & Doupe 2006) and mammalian (Dean et al 2005, 2006) auditory systems, in the visual cortex (Sharpee et al 2006) and in the somatosensory system (Maravall et al 2007). For a review, including the connections to information transmission, see Wark et al (2007). The retina offers an accessible model system in which to explore the mechanisms of such statistical adaptation, and these mechanisms seem quite rich and diverse (Rieke 2001, Kim & Rieke 2001, Baccus & Meister 2002, Kim & Rieke 2003). Most recent things from Fairhall & Rieke on the time constants of adaptation in retina. Look at recent work showing these effects in response to current injection in single neurons. Also say something about whether it is all too fast to be "real" adaptation, and whether this matters? For ideas about the adaptation/evolution of biochemical and genetic networks, see Francois et al (2007) and Francois & Siggia (2010).

Baccus & Meister 2002: Fast and slow contrast adaptation in retinal circuitry. SA Baccus & M Meister, *Neuron* **36,** 900–919 (2002).

Dean et al 2005: Neural population coding of sound level adapts to stimulus statistics. I Dean, NS Harper & D McAlpine, *Nature Neurosci* **8,** 1684–1689 (2005).

Dean et al 2006: Rapid neural adaptation to sound level statistics. I Dean, BL Robinson, NS Harper & D McAlpine, *J Neurosci* **28,** 6430–6438 (2006).

Francois et al 2007: A case study of evolutionary computation of biochemical adaptation. P Francois, V Hakim & ED Siggia, *Mol Syst Biol* **3,** 154 (2007).

Francois & Siggia 2010: Predicting embryonic patterning using mutual entropy fitness and in silico evolution. P Francois & ED Siggia, *Development* **137,** 2385–2395. (2010)

Kim & Rieke 2001: Temporal contrast adaptation in the input and output signals of salamander retinal ganglion cells. KJ Kim & F Rieke, *J Neurosci* **21,** 287–299 (2001).

Kim & Rieke 2003: Slow Na^+ inactivation and variance adaptation in salamander retinal ganglion cells. KJ Kim & F Rieke, *J Neurosci* **23,** 1506–1516 (2003).

Maravall et al 2007: Shifts in coding properties and maintenance of information transmission during adaptation in barrel cortex. M Maravall, RS Petersen, AL Fairhall, E Arabzadeh & ME Diamond, *PLoS Biology* **5,** e19 (2007).

Nagel & Doupe 2006: Temporal processing and adaptation in the songbird auditory forebrain. KI Nagel & AJ Doupe, *Neuron* **51,** 845–859 (2006).

Rieke 2001: Temporal contrast adaptation in salamander bipolar cells. F Rieke, *J Neurosci* **21,** 9445–9454 (2001).

Sharpee et al 2006: Adaptive filtering enhances information transmission in visual cortex. TO Sharpee, H Sugihara, AV Kurgansky, SP Rebik, MP Stryker & KD Miller, *Nature* **439,** 936–942 (2006).

Wark et al 2007: Sensory adaptation. B Wark, BN Lundstrom & A Fairhall, *Curr Opin Neurobiol* **17,** 423–429 (2007).

The ideas about efficient coding in the population of retinal ganglion cells go back to Barlow (1959, 1961), in delightfully original papers that I still find very inspiring. The precise mathematical formulation of these ideas took until Atick & Redlich (1990); van Hateren (1992) worked out essentially the same principles with invertebrate rather than vertebrate retinas in mind, and there are important precursors in the work of Srinivasan et al (1982) and Snyder et al on compound eye design (cited in Section I.A). The possibility that opponent color processing is an example of efficient coding was suggested by Buchsbaum & Gottschalk (1983), and the full analysis based on hyperspectral images is due to Ruderman et al (1998). An attempt to derive gene regulatory networks that optimize information transmission is described Tkačik et al (2009) and Walczak et al (2010).

Atick & Redlich 1990: Toward a theory of early visual processing. JJ Atick & AN Redlich, *Neural Comp* **2,** 308–320 (1990).

Barlow 1959: Sensory mechanisms, the reduction of redundancy, and intelligence. HB Barlow, in *Proceedings of the Symposium on the Mechanization of Thought Processes, volume 2*, DV Blake & AM Utlley, eds, pp 537–574 (HM Stationery Office, London, 1959).

Barlow 1961: Possible principles underlying the transformation of sensory messages. HB Barlow, in *Sensory Communication*, W Rosenblith, ed, pp 217–234 (MIT Press, Cambridge, 1961).

Buchsbaum & Gottschalk 1983: Trichromacy, opponent colour coding and optimum colour information transmission in the retina. G Buchsbaum & A Gottschalk, *Proc R Soc Lond B* **220,** 89–113 (1983).

van Hateren 1992: Real and optimal neural images in early vision. JH van Hateren, *Nature* **360,** 68–70 (1992).

Ruderman et al 1998: Statistics of cone responses to natural images: Implications for visual coding. DL Ruderman, TW Cronin & C–C Chiao, *J Opt Soc Am A* **15,** 2036–2045 (1998).

Srinivasan et al 1982: Predictive coding: A fresh view of inhibition in the retina. MV Srinivasan, SB Laughlin & A Dubs, *Proc R Soc Lond B* **216,** 427–459 (1982).

Tkačik et al 2009: Optimizing information flow in small genetic networks. I. G Tkačik, AM Walczak & W Bialek, *Phys Rev E* **80,** 031920 (2009).

Walczak et al 2010: Optimizing information flow in small genetic networks. II: Feed–forward interaction. AM Walczak, G Tkačik & W Bialek, *Phys Rev E* **81,** 041905 (2010).

Perhaps the most obvious evidence that the energy costs of spiking are significant is the fact that functional magnetic resonance imaging (fMRI) of the brian actually works: what one "sees" in these experiments are the changes in blood oxygenation that reflect the metabolic load associated with neural activity (Ogawa

et al 1990, 1992). The importance of fMRI has been one of the stimuli for a more detailed accounting of the energy budget of the brain (Attwell & Laughlin 2001, Raichle & Gusnard 2002). The description of spare/efficienct coding in the auditory system is based on the work of Lewicki (2002) and colleagues (Smith & Lewicki 2005, 2006). This grows out of earlier ideas of Lewicki & Sejnowksi (2000). For an overview of sparse coding and spikes, see Olshausen (2002) [more detailed refs to these ideas in the context of visual cortex].

Atwell & Laughlin 2001: An energy budget for signaling in the grey matter of the brain. D Atwell & SB Laughlin, *J Cereb Blood Flow & Metab* **21,** 1133–1145 (2001).

Lewicki 2002: Efficient coding of natural sounds. MS Lewicki, *Nature Neurosci* **5,** 356–363 (2002).

Lewicki & Sejnowski 2000: Learning overcomplete representations. MS Lewicki & TJ Sejnowski, *Neural Comp* **12,** 337–365 (2000).

Ogawa et al 1990: Brain magnetic resonance imaging with contrast dependent on blood oxygenation. S Ogawa, TM Lee, AR Kay & DW Tank, *Proc Nat'l Acad Sci (USA)* **87,** 9868–9872 (1990).

Ogawa et al 1992: Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. S Ogawa, DW Tank, R Menon, JM Ellerman, SG Kim, H Merkle & K Ugurbil, *Proc Nat'l Acad Sci (USA)* **89,** 5951–5955 (1992).

Olshausen 2002: Sparse codes and spikes. BA Olshauen, *Probabilistic models of the brain: Perception and neural function*, RPN Rao BA Olshausen & MS Lewicki, eds. pp 257–272 (MIT Press, Cambridge, 2002).

Raichle & Gusnard 2002: Appraising the brain's energy budget. *Proc Nat'l Acad Sci (USA)* **99,** 10237–10239 (2002).

Smith & Lewicki 2005: Efficient coding of time–relative structure using spikes. EC Smith & MS Lewicki, *Neural Comp* **17,** 19–45 (2005).

Smith & Lewicki 2006: Efficient auditory coding. EC Smith & MS Lewicki, *Nature* **439,** 978–982 (2006).

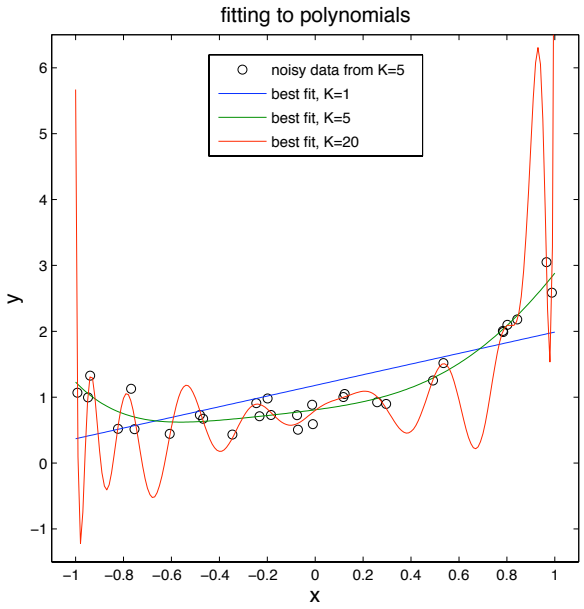

FIG. 161 Fitting to polynomials. We have a collection of data points (black circles), $\{x_n, y_n\}$, and we try to fit these data with polynomials of different degree K, with $K = 1, 5, 20$. We see that as the degree of the polynomial—what we think of intuitively as the complexity of our model—increases, we can get closer to the data points, but at the same time we are introducing wild fluctuations which seem unlikely to be correct. In fact, $K = 5$ is the correct answer, since the data points were generated by choosing the x_n at random, evaluating some fixed fifth–order polynomial, and then adding noise. To claim that we understand how to learn, we have to find a principled way of convincing ourselves that it's better to keep the poorer fit with the simpler model.

D. Gathering information and making models

The world around us, thankfully, is a rather structured place. Whether we are doing a careful experiment in the laboratory or taking a walk through the woods, the signals that arrive at our brains are far from random noise; there seem to be some underlying regularities or rules. Surely one task that all organisms must face is the learning or extraction of these rules and regularities, making models of the world, either explicitly or implicitly. In this section, we will explore how learning and making models is related to the general problem of efficient representation.

Perhaps the simplest example of learning a rule is fitting a function to data—we believe in advance that the rule belongs to a class of possible rules that can be parameterized, and as we collect data we learn the values of the parameters. This is something we all learned about in our physics lab classes (see Fig 161 for a reminder), and even this simple example introduces us to many deep issues. First, data usually come with some level of noise, and because of this any model really is (at least implicitly) a model of the probability distribution out of which the data are being drawn, rather than just a functional relationship. Indeed, one could argue that the general problem is always the problem of learning such distributions, and any rigid or deterministic rules emerge as a limit in which the noise becomes small or is beaten down by a large number of observations. The second point is that we would like to compare different models, often with different numbers of parameters. We have an intuition that simpler models are better, and we want to make this intuition precise—is it just a subjective preference, or is the search for simplicity something we can ground in more basic principles? A related point is that where the classical curve fitting exercises involve models with a limited number of parameters, we might want to go beyond this restriction and consider the possibility that the data are described by functions that are merely 'smooth' to some degree. Finally, we would like to quan-

tify how much we are learning—and how much *can* be learned—about the underlying rules given a limited set of data. If there are limits to how much we can learn, is it possible that biology has constructed learning machines which are efficient in some absolute sense, pushing up against these limits? So, let's plunge in ...

Imagine that we observe two streams of data x and y, or equivalently a stream of pairs (x_1, y_1), (x_2, y_2), $\cdots$, (x_N, y_N). Assume that we know in advance that the x's are drawn independently and at random from a distribution $P(x)$, while the y's are noisy versions of some function acting on x,

$$y_n = f(x_n; \boldsymbol{\alpha}) + \eta_n, \tag{862}$$

where $f(x; \boldsymbol{\alpha})$ is one function from a class of functions parameterized by $\boldsymbol{\alpha} \equiv \{\alpha_1, \cdots, \alpha_K\}$ and η_n is noise, which for simplicity we will assume is Gaussian with known variance σ^2. We can even start with a *very* simple case, where the function class is just a linear combination of basis functions, so that

$$f(x; \boldsymbol{\alpha}) = \sum_{\mu=1}^{K} \alpha_\mu \phi_\mu(x). \tag{863}$$

The usual problem is to estimate, from N pairs $\{x_i, y_i\}$, the values of the parameters $\boldsymbol{\alpha}$; in favorable cases such as this we might even be able to find an effective regression formula. Probably you were taught that the way to do this is to compute χ^2,

$$\chi^2 = \sum_n \left| y_n - f(x_n; \boldsymbol{\alpha}) \right|^2, \tag{864}$$

and then minimize to find the correct parameters $\boldsymbol{\alpha}$. You may or may not have been taught *why* this is the right thing to do, and this is what we would like to understand here.

If we assume that our model, Eq (862), is correct, what is the probability that we observe the data points $\{x_n, y_n\}$? Let's start by asking about the locations of the points x_n where we get samples of the functional relationship between x and y. In the standard examples of curve fitting, the examples are given to us and there is nothing more to say; thus, we might as well assume that the points x_n are chosen randomly and independently out of some distribution $P(x)$, perhaps just the uniform distribution on some interval. One might ask if there is a good choice for the next x_{n+1}, perhaps a point that will give us the maximal information about the underlying parameters $\boldsymbol{\alpha}$. This is the problem faced in the design of experiments—how do we choose what to measure given what we already know?—but let's leave this aside for the moment.

If we assume that the points $\{x_n\}$ are chosen out of some distribution, then conveniently our model in Eq (862) is a statement about the conditional probability distribution of y_n given x_n. Specifically, y_n is a Gaussian random variable with a mean value of $f(x_n; \boldsymbol{\alpha})$ and a variance of σ^2, so that

$$P(y_n | x_n, \boldsymbol{\alpha}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_n - f(x_n; \boldsymbol{\alpha}))^2}{2\sigma^2} \right]. \tag{865}$$

By hypothesis, the noise on every point is independent, which means that

$$P(\{y_n\} | \{x_n\}, \boldsymbol{\alpha}) = \prod_{n=1}^{N} P(y_n | x_n, \boldsymbol{\alpha}). \tag{866}$$

Now we can put things together to write the probability of the data given the parameters of the underlying model,

$$P(\{x_n, y_n\} | \boldsymbol{\alpha}) = \left[\prod_{n=1}^{N} P(y_n | x_n, \boldsymbol{\alpha}) \right] \times \left[\prod_n P(x_n) \right] \tag{867}$$

$$= \left[\prod_n P(x_n) \right] \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_n - f(x_n; \boldsymbol{\alpha}))^2}{2\sigma^2} \right] \tag{868}$$

$$= \exp\left[\sum_{n=1}^{N} \ln P(x_n) - \frac{N}{2} \ln(2\pi\sigma^2) - \frac{\chi^2}{2\sigma^2} \right] \tag{869}$$

where we identify χ^2 from Eq (864). Notice that the only place where the parameters appear is in χ^2, and $P \propto e^{-\chi^2/2\sigma^2}$. Thus, finding parameters which minimize χ^2 also serves to maximize the probability that our model could have given rise to the data. This sounds like a good thing to do, and certainly maximizing the probability of the data (usually called "maximum likelihood") feels more fundamental than minimizing χ^2. But what are we

really accomplishing by maximizing P?

We recall from Section IV.A that the entropy is the expectation value of $-\log P$, and that it is possible to encode signals so that the amount of "space" required to specify each signal uniquely is on average equal to the entropy. In such optimal encodings, each possible signal s drawn from $P(s)$ can be encoded in a space of $-\log_2 P(s)$ bits. Thus, any model probability distribution implicitly defines a scheme for coding signals that are drawn from that distribution, so if we make sure that our data have high probability in the distribution (small values of $-\log P$) then we also are making sure that our code or representation of these data is compact. What this means is that good old fashioned curve fitting really is all about finding efficient representations of data, which is the same principle that we discussed in the previous section in contexts ranging from the regulation of gene expression to neural coding. To be clear, in the earlier discussion we took for granted some physical or resource constraint (e.g., the noise level or limited number of molecules) and tried to transmit as much information as possible. Here we do the problem the other way, searching for a representation of the data that will require the minimum set of resources.

If we follow this notion of efficient representation a little further we can do better than just maximizing χ^2. The claim that a model provides a code for the data is not complete, because at some point we have to represent our knowledge of the model itself. One idea is to do this explicitly—estimate how accurately you know each of the parameters, and then count how many bits you'll need to write down the parameters to that accuracy and add this to the length of your code. Another idea is more implicit—you don't really know the parameters, all you do is estimate them from the data, so it's not so obvious that you should separate coding the data from coding the parameters, although this might emerge as an approximation. In this view what we should do is to integrate over all possible values of the parameters, weighted by some prior knowledge, and thus compute the probability that our data could have arisen from the *class* of models we are considering.

To carry out this program of computing the total probability of the data given the model class we need to do the integral

$$P(\{x_i, y_i\}|\text{class}) = \int d^K\alpha\, P(\boldsymbol{\alpha}) P[\{x_i, y_i\}|\boldsymbol{\alpha}] \tag{870}$$

$$= \int d^K\alpha\, P(\boldsymbol{\alpha}) \exp\left[-\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\chi^2(\boldsymbol{\alpha}; \{x_i, y_i\})\right]\left[\prod_n P(x_n)\right], \tag{871}$$

where $P(\boldsymbol{\alpha})$ is the a priori distribution of parameters, maybe just a uniform distribution on some bounded region. But remember that χ^2 as we have defined it is a sum over data points, which means it (typically) will be proportional to N. Thus, at large N we are doing an integral in which the exponential has terms proportional to N—and so we should use a saddle point approximation. To implement this approximation let's write

$$P(\{x_i, y_i\}|\text{class}) = \exp\left[-\frac{N}{2}\ln(2\pi\sigma^2)\right]\left[\prod_n P(x_n)\right]\int d^K\alpha\, e^{-Nf(\boldsymbol{\alpha})}, \tag{872}$$

where the effective "energy per data point" is

$$f(\boldsymbol{\alpha}) = \frac{1}{2N\sigma^2}\chi^2(\boldsymbol{\alpha}; \{x_i, y_i\}) - \frac{1}{N}\ln P(\boldsymbol{\alpha}) \tag{873}$$

The saddle point approximation is that

$$\int d^K\alpha\, e^{-Nf(\boldsymbol{\alpha})} \approx e^{-Nf(\boldsymbol{\alpha}^*)}(2\pi)^{K/2}\exp\left[-\frac{1}{2}\ln\det(N\mathcal{H})\right], \tag{874}$$

where $\boldsymbol{\alpha}^*$ is the value of $\boldsymbol{\alpha}$ at which $f(\boldsymbol{\alpha})$ minimized, and the Hessian $\mathcal{H}$ is the matrix of second derivatives of f at this point,

$$\mathcal{H}_{\mu\nu} = \left.\frac{\partial^2 f(\boldsymbol{\alpha})}{\partial\alpha_\mu\partial\alpha_\nu}\right|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}^*}. \tag{875}$$

At large N, $f(\boldsymbol{\alpha})$ is dominated by χ^2, so $\boldsymbol{\alpha}^*$ must be close to the point where χ^2 is minimized. Putting the pieces together, we have

$$-\ln P(\{x_i, y_i\}|\text{class}) \approx \frac{N}{2}\ln(2\pi\sigma^2) - \sum_{i=1}^{N}\ln P(x_i) + \frac{\chi^2_{\rm min}}{2\sigma^2} + \ln P(\boldsymbol{\alpha}^*) - \frac{K}{2}\ln 2\pi\sigma^2 + \frac{1}{2}\ln\det(N\mathcal{H})\,. \tag{876}$$

Note that $\mathcal{H}$ is a $K \times K$ matrix, and so $\det(N\mathcal{H}) = N^K \det(\mathcal{H})$. This allow us to group together terms based on their N dependence,

$$-\ln P(\{x_i, y_i\}|\text{class}) \approx -\sum_{i=1}^{N}\ln P(x_i) + \frac{\chi^2_{\rm min}}{2\sigma^2} + \frac{N}{2}\ln(2\pi\sigma^2) + \frac{K}{2}\ln N + \cdots\,, \tag{877}$$

where the first three terms are $\propto N$, and the terms $\cdots$ (including things we have neglected in the saddle point approximation) are constant or decreasing as $N \to \infty$. Again, the negative log probability measures the length of the shortest code for $\{x_i, y_i\}$ that can be generated given the class of models.

In Equation (877), the first term averages to N times the entropy of the distribution $P(x)$, which makes sense since by hypothesis the x's are being chosen at random. The second and third terms are as before, the length of the code required to describe the deviations of the data from the predictions of the best fit model; this also grows in proportion to N. The fourth term must be related to coding our knowledge of the model itself, since it is proportional to the number of parameters. We can understand the $(1/2)\ln N$ because each parameter is determined to an accuracy of $\sim 1/\sqrt{N}$, as in Fig 162, so if we start with a parameter space of size ~ 1 there is a reduction in volume by a factor of $\sqrt{N}$ and hence a decrease in entropy (gain in information) by $(1/2)\ln N$. Finally, the terms $\cdots$ don't grow with N.

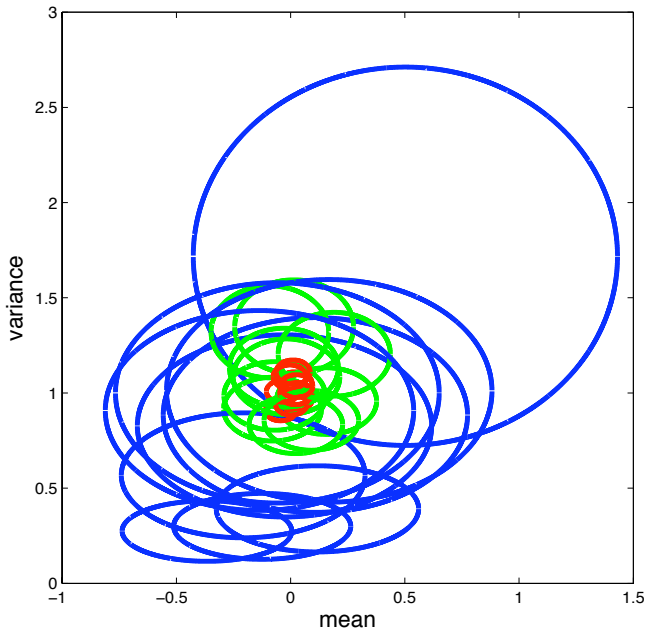

FIG. 162 Confidence limits on the estimation of mean and variance for a Gaussian distribution. In several independent experiments, we choose $N = 10$ (blue), $N = 100$ (green), or $N = 1000$ points out of a Gaussian distribution with zero mean and unit variance. We estimate the mean and variance from the data in the usual way, and draw error ellipses on the parameters that should contain 95% of the weight. We see that the linear dimensions of these ellipses shrink by $\sim 1/\sqrt{10}$ as N increase by a factor of 10. The (log) area inside the ellipses measures the entropy of our uncertainty in parameters, and decreases in this area correspond to gains in information.

Problem 165: Deriving the code length in a class of models. Fill in the details leading to Eq (877). Find an explicit form for the terms $\cdots$, and show that they do not grow with N. What assumptions do you need to make about the prior distribution $P(\boldsymbol{\alpha})$ in order to make this work?

What is crucial about the term $(K/2)\ln N$ is that it depends explicitly on the number of parameters. In general we expect that by considering models with more parameters we can get a better fit to the data, which means that χ^2 can be reduced by considering more complex model classes. But we know intuitively that this has to stop—we don't want to use arbitrarily complex models, even if they do provide a good fit to what we have seen. It is attractive, then, that if we look for the shortest code which can be generated by a class of models, there is an implicit penalty or coding cost for increased complexity. It is interesting from a physicist's point of view that this term emerges essentially from consideration of phase space or volumes in model space. It thus is an entropy–like quantity in its own right, and the selection of the best model class could be thought of as a tradeoff between this entropy and the "energy" measured by χ^2, a view to which we return below.

Thus Eq (877) tells that we have a *natural* penalty for the complexity of our model. While this term is linear in the number of parameters, it is only logarithmic in the number of data points. In contrast, $\chi^2_{\rm min}$ decreases with the number of parameters and is linear in the num-

ber of data point. In this way, the penalty for complexity becomes (relatively) less important the more data we gather: if we have only a few data points then although we could lower χ^2 by fitting every wiggle, the phase space factor pushes us away from this solution toward simpler models; if, however, the wiggles are consistent as we collect more data, then this factor becomes less important and we can move to the more complex models.

To see that these words really correspond to a quantitative theory, we have to generate a data set and go through the process of fitting via minimization of the 'code length' in Eq (877). For simplicity let's consider polynomial functions. We can pick a polynomial by choosing coefficients a_μ at random, say in the interval $-1 < a < 1$, where

$$f(x) = \sum_{\mu=0}^{K_{\rm true}} a_\mu x^\mu. \tag{878}$$

We'll confine our attention to the range $-5 < x < 5$; in this range the function $f(x)$ has some overall dynamic range (measured, for example, by its variance over this interval), and we'll assume the noise variance σ^2 is one percent of this 'signal' variance. Then we can generate points according to

$$y_{\rm n} = f(x_{\rm n}) + \eta_{\rm n}, \tag{879}$$

and try to fit. Fitting to any polynomial of degree K by minimizing χ^2 is a standard exercise, and in this way we find $\chi^2_{\rm min}(K)$. Then we can find the value of K that minimizes the total code length in Eq (877); this last step is just a competition between $\chi^2_{\rm min}(K)/\sigma^2$ and $(K+1)\ln N$. The results of this exercise are shown in Fig 163.

What we see in Fig 163 is that our qualitative description of the competition between complexity and goodness of fit really works. First we note that with a large number of data points, minimizing the code length zeroes in on the correct order of the underlying polynomial ($K \to K_{\rm true}$), despite the presence of noise that one could 'fit' using more complex models. Next, we see that for smaller numbers of data points, the shortest code is biased toward simpler models. In the limit that we only have a handful of data points, the shortest code is often a straight line ($K = 1$). Put another way, we start with a bias toward simple models, and only as we uncover more data can we support the adding of greater complexity.

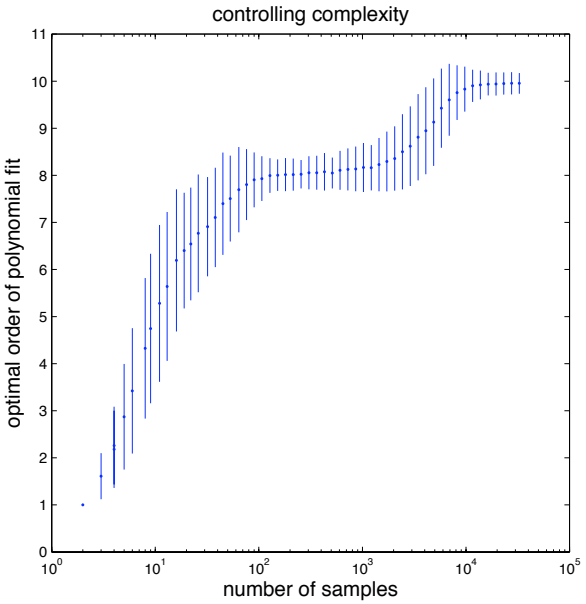

FIG. 163 Fitting to polynomials, part two. Choose the coefficients of a polynomial with degree 10 at random, and then choose points at random in the interval $-5 < x < 5$; there is added noise [as in Eq (862)] with a standard deviation set to be 1/10 of the overall dynamic range of the function $f(x)$. We then try to fit polynomials of order K, and find the value of K that minimizes the 'code length' in Eq (877). Since the result depends both on the particular value of the polynomial coefficients and on the particular points $x_{\rm n}$ that we happen to sample, we choose 500 examples and look at the mean (points) and standard deviation (error bars) across this ensemble of examples. Although the optimal order of the polynomial in any given example is, of course, an integer, fractional values arise from averaging over many examples.

Problem 166: Fitting and complexity. Generate a version of Fig 163 for yourself, doing a simulation which follows the steps outlined in the text. If you do this in MATLAB, you'll find the command `polyfit` to be useful. Some things to keep in mind:

(a.) Start with a small version of the problem, e.g. fitting to $N = 20$ data points.

(b.) Plot some of your intermediate results, just to get a feeling for what is going on. In particular, plot $\chi^2_{\rm min}$ as a function of K, verifying that higher order polynomials always give "better" fits in the sense of smaller χ^2.

(c.) Notice that $\chi^2_{\rm min}$ is a function of K and N, but also a function of the particular points $\{x_{\rm i}, y_{\rm i}\}$ you have "observed" in the experiment and of the particular parameters $\{a_\mu\}$ that specify the real function you are trying to learn. When you choose a different set of parameters and test points $\{x_{\rm i}\}$, from the same distribution, how different is the minimum "energy per data point" $\epsilon_{\rm min} = \chi^2_{\rm min}/(N\sigma^2)$ as a function of K? What happens to this variability as N gets larger?

(d.) Perhaps the most important thing is to verify that minimizing the code length really does control the complexity of the fit, selecting a nontrivial optimum K. Convince yourself that, as in Fig 163, the optimal K is small than $K_{\rm true}$ for small data sets, and approaches $K_{\rm true}$ as you analyze larger data sets.

There are many reasons to prefer simpler models, and certainly the idea that we entertain more complex models only as we collect more data is in accord with our sense of how we understand the world. But all of this can seem a little soft and squishy. Indeed, given the evident complexity of life and the world around us, one might start

to suspect that the preference for simple models is not an objective principle, but rather a subjective choice made by humans—and more often by scientists than by humans in general.[89] Even some technical discussions leave this impression of subjectivity, suggesting that while there must be a tradeoff between goodness of fit and complexity, the structure of this tradeoff is something that we are free to choose, perhaps inventing a new "penalty for complexity" tuned to the details of each problem. As physicists we are raised to be suspicious of overly complex models, but again this preference for simplicity is often couched in (surprisingly) soft words about the elegance or brevity of the equations that describe the model. What we have seen here is that all of this can be made much more precise.

The power of information theory in this context is that, by consistently measuring code lengths in bits we don't have to discuss our 'preference for simplicity' as a separate principle from goodness of fit. Deviations from the model (badness of fit) and the complexity of the model both add bits to the overall code length, and the relative contributions are calculable with no adjustable constants. The absence of unknown constants is important, since if we had to specify weights for the different terms we would once again inject subjectivity into the discussion of just how much we care about simplicity. Instead, we have one principle (search for the most compact description) and everything else follows. In particular, what follows is that limited experience (small N) biases us toward simpler models, while as we accumulate more experiences (ultimately, as $N \to \infty$) we can admit more complex descriptions of the world.

This is a very satisfying picture, and I am inclined to say that we can declare victory—we understand what we are doing when we make models, why simple models are preferable, and how the support for more complex models emerges. Nonetheless, there are several loose ends, and I'm not sure that I know how to tie them all up.

The first and most obvious problem is that our discussion makes sense as long as we specify in advance a class of models, and more seriously a hierarchy of such classes with increasing complexity. It's not at all obvious how to do this. Worse yet, plausible but wrong ways of doing this can lead to weird results, for example if we have a function well described by a Fourier series with just a few terms, but we try fitting polynomials. Simplicity and complexity have meaning as code lengths only if we have a defined ensemble of possibilities to choose from, in much the same way that Shannon's original discussion of the information gained on hearing the answer to a question (Section IV.A) starts with the assumption that we know the distribution out of which answers will be drawn.

A second, and perhaps related, problem is that we are discussing models with a finite number of parameters. It might seem more natural, for example, to imagine that the relationship between x and y is just some smooth function, not necessarily describable with a finite number of parameters; that is, $f(x)$ should live in a function space and not in a finite dimensional vector space. Now we have to specify a prior distribution not on the parameters, as with $P(\boldsymbol{\alpha})$ above, but on the functions themselves $P[f(x)]$. The simplest version of this problem is not with functional relations but just with probability distributions: suppose that we observe a set of points $x_1, x_2, \cdots, x_N$, which we assume are drawn randomly and independently out of a distribution $Q(x)$; how do we estimate Q? If the distribution we are looking for belongs to a family with a finite number of parameters, we proceed as before, but if all we know is that $Q(x)$ is a smooth function then we have to specify a prior probability distribution on this space of distributions. From a physicist's point of view, probability distributions on such function spaces are just scalar field theories, and one can carry a fair bit of technology over to do real computations. The lesson from these computations is that, with some reasonable priors to implement what we mean by "smooth," everything works as it does in the case of finite parameters, but the prior does matter.

Problem 167: Taming the singularities. The basic problem in trying to learn a continuous probability distribution is to explain why, having observed a set of points $x_1, x_2, \cdots, x_N$, we shouldn't just guess that the distribution is of the form

$$Q(x) \sim \frac{1}{N} \sum_{\mathrm{i}=1}^{N} \delta(x - x_{\mathrm{i}}), \tag{880}$$

which of course generates precisely the data we have observed with maximal (infinite!) probability density. We all know that this is the wrong answer, and the role of priors on the space of distributions is to express this knowledge. A very different approach to taming the singularities is sometimes called Kernel density estimation, in which we search for a probability distribution in the form

$$Q(x) = \frac{1}{K} \sum_{\mathrm{j}=1}^{K} \frac{1}{\ell} F\left(\frac{x - y_{\mathrm{j}}}{\ell} \right), \tag{881}$$

where ℓ is again a characteristic length scale, $F(z)$ is some 'blob–like' function, and the y_{j} are the centers of the blobs; F is normalized so that $\int dz\, F(z) = 1$. For concreteness let

$$F(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}. \tag{882}$$

If we let $K = N$ (generally not such a good idea), then it should be clear that the model which generates the data with the highest probability is one in which the kernel centers are on top of the data points, $y_{\mathrm{i}} = x_{\mathrm{i}}$ for all i. It should also be clear that this probability of the data increases for smaller ℓ, diverging as $\ell \to 0$. But we

[89] One could add that even among scientists, physicists have a special affinity for simple models, often to the point of being the punchline in jokes, as in " ... consider the case of the spherical horse."

know that, to get control over complexity, we should compute the *total* probability of generating the data in this class of model. In this case the parameters of the model are the kernel centers $\{y_i\}$. Assume that everything happens in a box, so that $0 < x < L$, and and similarly for $\{y_i\}$; by translation invariance the prior on the ys should be flat in this box. Calculate the total probability that this class of models generates the data in the limit $\ell \to 0$. Is the answer finite? If so, this means that the phase space factors are just strong enough to compensate for the 'goodness of fit' and prevent anything from diverging in this limit. Can you find any other approximations that allow you to say anything about the optimal value of ℓ?

Quite generally, when we compute the total probability that a model can generate data, we are doing integrals like

$$P(\{x_i\}|\text{model class}) = \int DQ\, P[Q(x)] \prod_{i=1}^{N} Q(x_i), \quad (883)$$

where $P[Q(x)]$ is the probability distribution function(al) on the space of distributions. It embodies all our prior knowledge, in whatever form—that the distribution can be described by a few parameters, or merely that it is smooth in some sense. To understand what is happening in this integral, it is useful to measure possible distributions $Q(x)$ relative to the true distribution $Q_{\text{true}}(x)$,

$$P(\{x_i\}|\text{model class}) = \left[\prod_{i=1}^{N} Q_{\text{true}}(x_i)\right] \int DQ\, P[Q(x)] \prod_{i=1}^{N} \left[\frac{Q(x_i)}{Q_{\text{true}}(x_i)}\right]. \quad (884)$$

We can collect the product into an exponential,

$$P(\{x_i\}|\text{model class}) = \left[\prod_{i=1}^{N} Q_{\text{true}}(x_i)\right] \int DQ\, P[Q(x)] \exp\left[N \frac{1}{N} \sum_{i=1}^{N} \ln\left(\frac{Q(x_i)}{Q_{\text{true}}(x_i)}\right)\right], \quad (885)$$

and we recognize that the average over data points x_i approaches, at large N, an average over the true distribution,

$$P(\{x_i\}|\text{model class}) \to \left[\prod_{i=1}^{N} Q_{\text{true}}(x_i)\right] \int DQ\, P[Q(x)] \exp\left[N \int dx\, Q_{\text{true}}(x) \ln\left(\frac{Q(x)}{Q_{\text{true}}(x)}\right)\right] \quad (886)$$

$$= \left[\prod_{i=1}^{N} Q_{\text{true}}(x_i)\right] \int d\epsilon\, \mathcal{N}(\epsilon) e^{-N\epsilon}, \quad (887)$$

where

$$\epsilon = \int dx\, Q_{\text{true}}(x) \ln\left(\frac{Q_{\text{true}}(x)}{Q(x)}\right) \quad (888)$$

is the Kullback–Leibler divergence between the distribution $Q(x)$ and the true distribution, and

$$\mathcal{N}(\epsilon) = \int DQ\, P[Q(x)] \delta\left[\epsilon - \int dx\, Q_{\text{true}}(x) \ln\left(\frac{Q_{\text{true}}(x)}{Q(x)}\right)\right] \quad (889)$$

counts the (weighted) volume in model space that is at KL divegence ϵ away from the right answer. Now ϵ, which is a "goodness of fit" between the model and the data, can be thought of as an energy, while the (log) volume in model space is an entropy, $\mathcal{N}(\epsilon) = e^{S(\epsilon)}$. If we imagine the the model space has a finite but large dimensionality K, then we expect that the entropy will be extensive, $S(\epsilon) = Ks(\epsilon)$. So, when the dust settles,

$$P(\{x_i\}|\text{model class}) \propto \int d\epsilon\, \exp\left[-N\left(\epsilon - \frac{K}{N} s(\epsilon)\right)\right]. \quad (890)$$

Thus, at large N, the integral is dominated by the minimum of the free energy density, $f = \epsilon - Ts(\epsilon)$, where the role of temperature is played by $T = K/N$. This calculation makes explicit the idea that learning really is statistical mechanics in the space of models, and that see-

ing more examples is like like lowering the temperature, 'cooling' the system into an ordered state around the right answer. Depending on space of possible models, and hence the function $s(\epsilon)$, there can be phase transitions—a sudden jump, as we collect more examples, from wandering around in model space to having a compelling fit to the data.

What would it mean to have a phase transition in learning? As we accumulate more examples, we are lowering the effective temperature in the equivalent statistical mechanics problem. At first this doesn't do very much, in the same way that lowering the temperature of water from 100 °C to 30 °C doesn't do very much much. But, at some point, a relatively small change in the number of examples we have seen produces a huge change in the distribution over models, freezing into a small volume surrounding the correct answer. This would be something like the subjective "aha!" experience, where we suddenly seem understand something or master a skill after a very period of experience or training. Although we have all (I hope) experienced this phenomenon, it is not so easy to study quantitatively, and so I think we have no idea whether the statistical mechanics approach to learning provides a useful guide to understanding this effect.

It is interesting to look at the history of studies in animal learning in the light of these results. Already in the 1920s and 30s it was clear that, at certain tasks, animals could exhibit "sudden" rather than gradual learning. Although this was well before Hebb, and decades before the observation of changes in synaptic strength driven by the correlation between pre– and post–synaptic neurons (see Section [point back to previous chapter; be sure it's there!]), there was a general view that learning relied upon statistical association, and thus should be a continuous process. Thus there was a question whether sudden learning represents a new mechanism, beyond associative processes. The mapping of learning onto a statistical mechanics problem reminds us that when there are many degrees of freedom, continuous dynamics can have nearly discontinuous consequences.

Before leaving the image of energy/entropy competition behind, we should note a caveat. In getting to Eq (890), we have first allowed N to become very large, so that averages over samples can be replaced by averages over the underlying distribution, and then used the resulting formulae with finite N to say something about how learning proceeds. Evidently this is dangerous. It also was controversial when it first emerged, since the results seemed to conflict with an approach by computer scientists which emphasized bounds on the learning curve. To explain how all this was resolved would take us far afield, so I'll point to the references at the end of this section. When the dust settles, there is a well defined approximation that leads to Eq (890), and the resulting predictions can be made rigorous and shown to be consistent with known bounds.

It would be good to connect these ideas with experiment. To what extent is our (or other animals') performance in situations where we learn understandable in terms of these theoretical structures? A big problem here is what to measure. In the examples discussed above, what is being learned is a probability distribution, or some set of parameters describing the data that we observe. It's not so easy to ask even a human subject to report on their current estimates of these parameters, and it's completely unclear how we would do this in simpler organisms. In practice, subjects are usually asked to make a decision; in classical work on pigeons the decision is to peck or not to peck at a target, and for humans subjects are simply asked a yes/no question, or asked to push one of a small set of buttons. Evidently the bandwidth of these experiments is limited—although we may be continuously updating an internal model with many parameters, what we report is on the order of one bit, yes or no.

One context that comes closer to the theoretical discussion, albeit in a simple form, concerns making decisions when the alternatives come with unequal probabilities. This harkens back to out earliest topic, a human observer waiting for a dim flash of light in a dark room. As we discussed in that context, optimal decisions, deciding that a signal is convincingly above the background of noise, are achieved by setting a threshold that depends on the probability that the signal is present [need a definite pointer]. If this probability can change over time, then it must be learned. More prosaically, if we have to choose between two alternatives even in a limit where they are fully distinguishable, but the rewards for the different choices vary probabilistically, then we have to learn something about the underlying probabilities of reward in order to develop a sensible strategy. These sorts of experiment have attracted interest because they might connect to our economic behavior, and because they provide settings in which we can search for the neural correlates of the subject's estimate of probability and value.[90]

There is a classical literature showing that human observers adjust their criteria for detecting signals to the probability that the signals occur. The question about learning is really how long it takes the subject to make this adjustment. In the simplest case, the probability changes suddenly, and we look for a change in behavior in response. If the only behavioral output is a decision among two alternatives, we as observers also need to go

[90] I think it is fair to say that the concept of "value" has attracted more attention in this context, because it seems more connected to economics. Indeed, there is now a whole field described as "neuro–economics." But perhaps the probabilistic nature of our inferences, even in the economic context, have been given less attention than they deserve.

through an inference process to decide when is the first sign of a response. In such an experiment, we have a complete probabilistic description of the trajectory taken by the sensory stimuli or rewards, so at any moment we can calculate the probability that the signals being shown are consistent with constant parameters or a recent, sudden change. Given the responses of the subject, we can also ask for the moment at which we see the first statistical sign of a change in behavior. In experiments where rats experience changing reward probabilities, the change in behavior occurs at times so soon after the changes in probability that the best evidence for the change is modest, corresponding to probabilities in the range 0.1 to 0.9; only rarely (in $\sim$ 20% of trials) do rats wait to reach 99% certainty. On these very short time scales, the rate at which the rat collects rewards changes very little, suggesting that changes in strategy really are driven by learning the underlying probabilities, rather than tinkering until rewards accumulate.

In a similar spirit, we can do a longer experiment, with the probabilities jumping among different levels, and track the dynamics of the behavior. [Have to decide how much to say here. Would like to connect with result from Corrado et al suggesting that filtering of experience to generate internal model of probability is near–optimal. Could do the calculation in a simple case, then point to

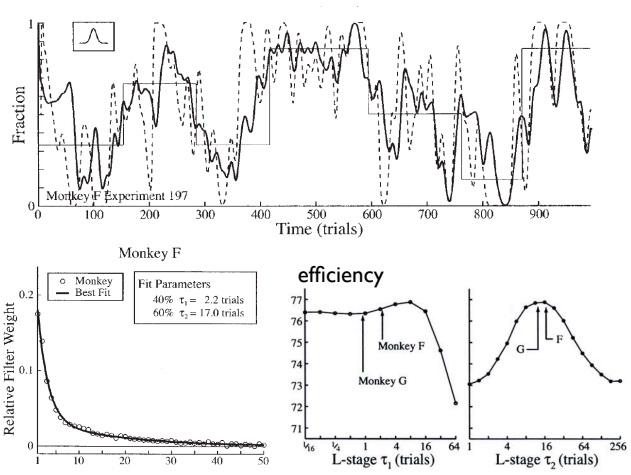

FIG. 164 Tracking the changing probabilities of reward, from Corrado et al (2005). At top, the local frequencies of choosing one of two alternatives (solid) and being rewarded (dashed), when the probability of this choice being rewarded jumps among different levels as shown (thin line); frequencies are computed from discrete events by smoothing with the Gaussian kernel shown in the inset. At bottom left, the filter inferred from the relationship between rewards and subsequent choices. At bottom right, the efficiency of collecting rewards averaged over the whole session, assuming that the subject implements a filter with the times constants as shown. The subjects' behaviors are best fit by parameters that generate efficiencies within one percent of the optimum.

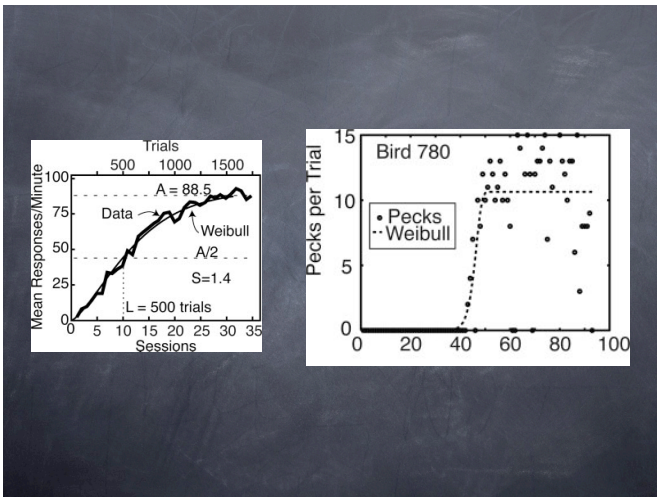

FIG. 165 Learning curves in individuals vs groups (Gallistel et al 2004). [Need to give a full explanation of the experiment!] At left, average performance in a large population of birds improves gradually and very slowly, requiring many hundreds of trials before reaching its half maximal level. At right, performance measured in one individual bird is noisy (because we have access only to the number of pecks as a behavioral output), but makes a relatively sudden transition to near saturating performance as the animal experiences $\sim$ 10 additional examples.

Fig 164 ... Need to digest the paper better, though.]

One approach to adding bandwidth to experiments on learning is to average over many subjects, so that the performance after N examples can be measured as a real number (e.g., the probability of subjects getting the right answer) even though the data from individuals is discrete (yes/no answers). But, as emphasized in Fig 165, this can be misleading. Individual subjects seem to learn simple tasks abruptly, but with transitions after different numbers of trials, so that average "learning curves" are smooth and gradual. This is interesting, because more abrupt learning reminds us of performance as a function of signal–to–noise ratio in discrimination tasks, and because theory along the lines described above often predicts relatively rapid learning when the space of possibilities is small. The variations across individuals may then reflect differences in how the 'small problem' posed by the particular experimental situation is weighted within the much larger set of possible behaviors available to the organism. But much needs to be done to make this precise.

Another approach to increasing the bandwidth of behavioral experiments is to look at continuous motor outputs rather than decisions. An example is if we have to move an object through a medium that generate an unknown, anisotropic mobility tensor; as we practice, we learn more about the parameters of our environment and can move more accurately. Importantly, each trial of such an experiment generates an entire movement trajectory

rather than just a single discrete decision. Analysis of these trajectories can reveal how the errors we make in one trial influence the change of our internal model on the next trial. [Should have a figure—maybe combine something from Shadmehr et al plus saccadic latency vs probability?] Although this emphasizes learning of parameters that influence the movement itself, the fact that some movements are made in extraordinary precise relations to sensory inputs (e.g., as we follow a moving target with our eyes), and that we can learn to anticipate the need for such movements (e.g., as targets follow predictable trajectories), suggests that analysis of continuous movements should more generally provide us with a path to examine more details of the brain's internal model of the world. A simple version of this idea is that the latency for us to move our eyes toward one of two suddenly appearing targets depends on the relative probabilities of the targets—we move more quickly toward targets of higher probability, as shown in Fig 166, and it is tempting to think that the latency of movement gives us a readout of the brain's estimate of this probability. Again, there is much to do here.

Thus far our examples of learning have been "passive." That is, the learner experiences a data stream from which inferences can be drawn, but there is no way for the learner to shape the data stream, selecting observations which might be especially informative. [Give a discussion of infotaxis. This is interesting both as an active learning problem and as an example where gathering information substitutes effectively for "goal–directed behavior.]

Finally, a theoretical point. We have emphasized that learning a model amounts to building an efficient representation of the data we have observed, and hence the "goal" of learning is no different than the goals proposed in the previous section for the transmission of information through neural or genetic networks. This theoretical unity is attractive. But one might worry—why do we care about representing what we have observed in the past? What matters, to follow the discussion at the end of Section IV.B, is what is of use in guiding our actions in the future. Thus, presumably we learn models that describe data collected in the past because we expect these models to still be true in the future, and this allows us to make successful predictions. How does this connect to our ideas about efficient representation?

We recall from Section IV.B that the predictive information in a time series, that is the information which observations on the past provide about the future, is equal to the subextensive component of the entropy. In the course of evaluating the probability of data given a class of models, in Eq (877), we have implicitly calculated this subextensive entropy. Specifically, we found that the negative log probability of a set of data at N time points had a term proportional to N (the extensive piece), and a term that grows only logarithmically with N (the leading subextensive piece), as in Eq [**]. Thus, when we are observing a time series from which we can learn a model with K parameters, there is a subextensive entropy and hence predictive information $\sim (K/2)\log_2 N$ bits. The "meaning" of this predictive information is precisely that we know something about the parameters underlying the data, and on the hypothesis that these parameters are constant we can predict something about the future.

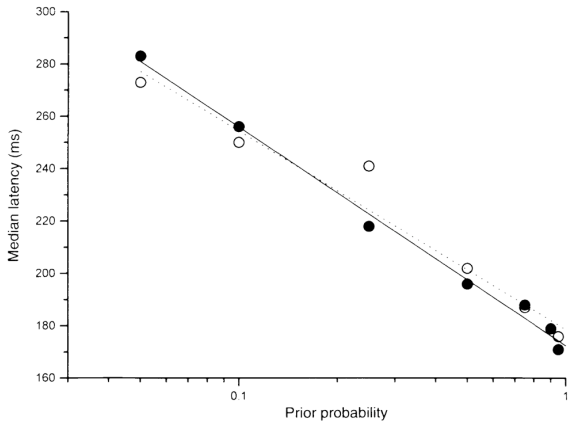

FIG. 166 Latency for saccadic eye movements to targets of varying probability, from Carpenter & Williams (1995). Subjects are asked to move their eyes to a target which appears at a random time after they fixate on a small spot. The target is either to the left or right, with varying probabilities in blocks of trials. For two subjects (filled and empty symbols), one collects all the trials in which the subject move to a target of probability p, and computes the mean latency of the eye movement. Do we want to say anything about the distribution of latencies?

Problem 168: Predictive information in learning. [One more problem with the details.]

When we observe N data points, the total amount of information we have collected is a number of bits proportional to N. But in the case we are considering, there are just $\sim (K/2)\log_2 N$ bits of information about the future. If we can separate these predictive bits from the nonpredictive background, we will have learned the parameters of the underlying model. Thus, compressing the data while preserving the predictive information is exactly the same problem as learning. Interestingly, if we live in a world described by a complex model (large K), then the amount of predictive information is much larger than the information needed to describe the present.

I think that our modern understanding of the preference for simple models, as explained here, is quite important, well known in certain circles, but less widely appreciated than it should be. Part of the difficulty is the presence of many independent threads in the literature. Rissanen had a very clear point of view which is essentially that presented here, although in different language; sources go back at least to Rissanen (1978), with a summary in Rissanen (1989). The problem became more urgent with the emergence of neural networks, which could be viewed as models with very large numbers of parameters. In this context, MacKay (1992) understood the critical role of 'Occam factors,' the integrals over parameter values that favor simpler models; see also his marvelous textbook (MacKay 2003). Balasubramanian (1997) generalized these ideas and translated them into physicists' language, showing how the Occam factors can be thought of as entropy in the space of models. Certainly I learned a lot from talking to Balasubramanian, and from working out these ideas in the context of a field theoretic approach to learning distributions (Bialek et al 1996). For the case of the spherical horse, see Devine & Cohen (1992).

Balasubramanian 1997: Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. V Balasubramanian, *Neural Comp* **9,** 349–368 (1997).

Bialek et al 1996: Field theories for learning probability distributions. W Bialek, CG Callan & SP Strong, *Phys Rev Lett* **77,** 4693–4697 (1996).

Devine & Cohen 1992: *Absolute Zero Gravity: Science Jokes, Quotes and Anecdotes.* B Devine & JE Cohen (Fireside Press, Philadelphia, 1992).

MacKay 1992: A practical Bayesian framework for backpropagation networks. DJC MacKay, *Neural Comp* **4,** 448–472 (1992).

MacKay 2003: *Information Theory, Inference, and Learning Algorithms.* DJC MacKay (Cambridge University Press, Cambridge, 2003).

Rissanen 1978: Modeling by shortest data description. J Rissanen, *Automatica* **14,** 465–471 (1978).

Rissanen 1989: *Stochastic Complexity and Statistical Inquiry* J Rissanen (World Scientific, Singapore, 1989).

The study of neural networks also led to a very explicit formulation of learning as a statistical mechanics problem (Levin et al 1990). Within this framework it was discovered that there could be phase transitions in the learning of large models (Seung et al 1992), and these can be understood as a competition between energy (goodness of the fit) and entropy in the space of models; the effective temperature is the inverse of the number of examples we have seen, so the system 'cools' as we collect more data. Meanwhile the computer scientists have developed approaches to learning rules and distributions that focused on rigorous bounds—given that we have seen N examples, can we guarantee that our inferences are within ϵ of the correct model with probability $1-\delta$? These ideas have their origins in Vapnik and Chernovonenkis (1971) and Valiant (1984). The rapproachment between the different approaches was given by Haussler et al (1996). For an early discussion about sudden vs. gradual learning, see Spence (1938). For a modern example, emphasizing the need for a unified approach to sudden and gradual learning, see Rubin et al (1997).

Haussler et al 1996: Rigorous learning curve bounds from statistical mechanics. D Haussler, M Kearns, HS Seung & N Tishby, *Machine Learning* **25,** 195–236 (1996).

Levin et al 1990: A statistical approach to learning and generalization in layered neural networks. E Levln, N Tishby & SA Solla, *Proc IEEE* **78,** 1568–1574 (1990).

Rubin et al 1997: Abrupt learning and retinal size specificity in illusory–contour perception. N Rubin, K Nakayama & R Shapley, *Curr Biol* **7,** 461–467 (1997).

Seung et al 1992: Statistical mechanics of learning from examples. HS Seung, H Sompolinsky & N Tishby, *Phys Rev A* **45,** 6056–6091 (1992).

Spence 1938: Gradual versus sudden solution of discrimination problems by chimpanzees. KW Spence, *J Comp Psych* **25,** 213–224 (1938).

Valiant 1984: A theory of the learnable. LG Valiant, *Commun ACM* **27,** 1134–1142 (1984).

Vapnik & Chernvonenkis 1971: On the uniform convergence of relative frequencies of events to their probabilities. VN Vapnik & AY Chervonenkis, *Theory of Probability and its Applications* **16,** 264–280 (1971).

Do we ned a pointer back to Green and Swets, or some other reference about changing criteria in relation to changing probabilities?

The analysis of the response to sudden changes in probability is by Gallistel et al (2001). The experiments on fluctuating probabilities with primate subjects are by Sugrue et al (2004), and the analysis in Fig 164 is by Corrado et al (2005). For views on the emerging ideas of neuro–economics, see Glimcher (2003) and Camerer et al (2005). Measurements on learning through trial–by–trial analysis of continuous movement trajectories were pioneered by Thoroughman & Shadmehr (2000), who considered human arm movements; they had a particular view of the class of models that subjects use in these experiments, which simplified their analysis, but the idea is much more general. For measurements on the precision of tracking eye movements, see Osborne et al (2005, 2007). For the beautiful relationship between latency and target probability in saccadic eye movements, see Carpenter & Williams (1995). The idea of "infotaxis" is due to Vergassola et al (2007). Regarding the connection between learning and predictive information, see Bialek et al (2001) in Section IV.B.

Camerer et al 2005: Neuroeconomics: How neuroscience can inform economics. C Camerer, G Lowenstein & D Prelec, *J Econ Lit* **43,** 9–64 (2005).

Carpenter & Williams 1995: Neural computation of log likelihood in control of saccadic eye movements. RHS Carpenter & MLL Williams, *Nature* **377,** 59–62 (1995).

Corrado et al 2005: Linear–nonlinear–Poisson models of primate choice dynamics. GS Corrado, LP Sugure, HS Seung & WT Newsome, *J Exp Anal Behav* **84,** 581–617 (2005).

Gallistel et al 2001: The rat approximates an ideal detector of changes in rates of reward: Implications for the law of effect. CR Gallistel, TA Mark, AP King & P Latham, *J Exp Psych: Animal Behav Proc* **27,** 354–372 (2001).

Gallistel et al 2004: The learning curve: Implications of a quantitative analysis. CR Gallistel, S Fairhurst & P Balsam, *Proc Nat'l Acad Sci (USA)* **101,** 13124–13131 (2004).

Glimcher 2003: *Decisions Uncertainty and the Brain: The Science of Neuroeconomics* PW Glimcher (MIT Press, Cambridge, 2003).

Osborne et al 2005: A sensory source for motor variation. LC Osborne, SG Lisberger & W Bialek, *Nature* **437,** 412–416 (2005).

Osborne et al 2007: Time course of precision in smooth pursuit eye movements of monkeys. LC Osborne, SS Hohl, W Bialek & SG Lisberger, *J Neurosci* **27,** 2987–2998 (2007).

Sugrue et al 2004: Matching behavior and the representation of value in the parietal cortex. LP Sugrue, GS Corrado & WT Newsome, *Science* **304,** 1782–1787 (2004).

Thoroughman & Shadmehr 2000: Learning of action through adaptive combination of motor primitives. KA Thoroughman & R Shadmehr, *Nature* **407,** 742–747 (2000).

Vergassola et al 2007: 'Infotaxis' as a strategy for searching without gradients. M Vergassola, E Villermaux & BI Shraiman, *Nature* **445,** 406–409 (2007).

E. Perspectives

Optimizing information transmission, or maximizing the efficiency with which information is represented, is the sort of abstract, general principle that physicists find appealing. At the same time, this abstraction makes us suspicious about its relevance to the nitty–gritty of life. Thus, while information is essential for survival, surely much of what organisms do is bound up in the fact that some bits are more useful than others, and in the challenges of acting rather than just collecting data. In this Chapter I have tried to show both how interesting predictions flow from the abstract principles, and how these principles connect, sometimes surprisingly, to the more quotidian facts of life. It is surely too early, in this as in any other section of the course, to decide if some candidate theoretical principles are "right," and in any case I am not a disinterested observer. What I would like to emphasize here is that thinking about the optimization of information transmission has been productive, not least because it suggests genuinely new kinds of experiments. In many systems, these experiments have generated interesting results, independent of the theoretical motivation. In many other systems, even the first generation of experiments remains to be done.

Perhaps the most important point about information theoretic optimization principles is that they force us to think about biological systems in context. Whereas classical biology routinely considered organisms in their natural setting, as biology has modernized and become more mechanistic, we see more and more work on systems shorn of their context. To give an example, it may be that the best studied example of the regulation of gene expression is the *lac* operon in *E. coli*. But how much do we know about the distribution of lactose concentrations encountered by these cells in their natural environments? We know that, under many conditions, the total number of *lac* repressor proteins in the cell is small, but what is the dynamic range of this number over the lifetime of the organism? Vastly more is known about the details of the DNA sequences that are targeted by transcription factors involved in the regulation of metabolic genes than is known about the real world variations in nutrient conditions that create the need for metabolic regulation.

In the case of neural information processing, the ethologists—who often study systems specialized for the processing of particular sense data, such as bird song or bat echolocation—provided a persistent reminder about the importance of the natural context in understanding biological function. Perhaps our human abilities to deal with a seemingly much wider range of data and tasks generated some resistance to thinking that lessons from a barn owl or an electric fish could be of relevance to how we explore higher brain function. The claim that at least some aspects of neural circuitry are arranged to generate efficient representations of incoming sense data provided a counterpoint, suggesting that even for a "general purpose" sensory system, context matters. By now there is a whole subfield of neuroscience focused on the structure and processing of natural signals, a field which we might think of as a modern, quantitative development of the early work in ethology. Because our sense organs are such high quality devices, there are substantial experimental challenges in characterizing their natural inputs and in delivering controlled versions of these natural signals in the laboratory. Precisely because natural signals are rich and complex, analyzing neural responses to these signals poses significant theoretical challenges (see, for example, Appendix A.7). Progress on these experimental and theoretical problems is giving us more powerful tools with which to explore the brain, again independent of the sometimes distant motivation from optimization principles.

Thinking about information flow encourages us to ask about the structure of natural behavioral outputs, as well as natural sensory inputs. In the attempt to quantify animal (and human) behavior in the laboratory, there has been a tradition of constraining this behavior to a small, discrete set of alternatives, and this has been enormously powerful, not least because such constrained experiments are amenable to analyses in terms of signals and noise as in our initial discussion of photon counting in vision. Similarly, experiments on the control of gene expression in single celled organisms often have focused on the "switch" in expression patterns associated with a sudden transition from one nutrient source to another. Even the ethologists tended to categorize, collapsing whole ranges of behavior onto a limited space of discrete choices. But behavior, from single cells to entire humans, is vastly richer than choosing among discrete alternatives. As the technology for monitoring behavior improves, it becomes possible to ask if the continuous variations in natural behaviors are just noise, or are related systematically to the goals and context. Even if behavior really is composed of choice among a small set of stereotyped possibilities—such as running and tumbling in *E. coli*—the timing of these choices can convey information about the sensory inputs that drive them.

We have the impression that we are bombarded by complex data, and that our behaviors are relatively limited. But the inputs to our sensory system are highly structured, presumably because they derive from a limited set of causes and effects in the environment, and hence carry much less information about what is really "out there" than one might guess from the available bandwidth; our receptors provide limited, noisy views of these inputs, reducing the information still further (see, for example, Problem [info in cone array]). At the other

end, our motor outputs in fact are quire rich, even if we tend to coarse grain and categorize these behaviors into limited classes. Could it be that motor outputs are so carefully shaped and timed in relation to sensory inputs from the environment that we (and other organisms) are making use of a large fraction of the information available about this environment? There is a huge experimental challenge in tracking information flow all the way from sensory input to motor output, even in simple cases, and in more complex cases there is a substantial theoretical challenge in providing a framework for the analysis of such data.

One of the most important aspects of information theory is the fact that bits have value. This is why, for example, there is a minimum number of bits we need to send over a telephone connection to be sure that speech is intelligible and speakers identifiable. For living organisms, the value of bits depends on many details, perhaps more detail than, as physicists, we would like to think about. What we can say, however, is that bits which have no predictive power are valueless, and that most of the bits we have collected over our lifetime are in this valueless category. Thus, separating predictive information from the background of non–predictive clutter is a formidable, and biologically relevant, challenge. Importantly, this very general task seems to contain within it, as special cases, problems ranging from signal processing to learning, problems that we usually think of as belonging to different levels of biological organization with very different mechanisms. Perhaps this is, after all, a path to the sort of general principle we are seeking.