Chapter 4

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. 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 section of the course 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. We'll see that these ideas have the potential to tie together phenomena ranging from the control of gene expression in bacteria to learning in the brain. We begin with the foundations.

4.1 Entropy and information

Section last updated November 25, 2008.

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

$$S = -\sum_{\mathrm{n}} p_{\mathrm{n}} \log_2 p_{\mathrm{n}}\,\mathrm{bits}. \tag{4.1}$$

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 [Shannon 1948].[1]

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, and is something we should revisit. If we leave out the test of the questions and answers themselves, then all that remains are the probabilities p_{n} of hearing the different answers, and so Shannon assumes that the information gained must be a function of these probabilities, $I(\{p_{\mathrm{n}}\})$. The challenge is to determine this function.[2]

[1] To a remarkable extent, Shannon's original work provides a complete and accessible guide to the foundations of the subject. 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 TM Cover & JA Thomas, *Elements of Information Theory* (Wiley, New York, 1991); there is also a second edition (2006).

[2] 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 previously unimaginable answer to our question. We will see (once we have built some structure on which to stand) that there are problems and controversies about the use of information in the Shannon sense to describe the complexities of biological systems. I think that some of these problems really come from this basic assumption that there is a measure of information that transcends the details of the particular questions and answers one encounters in

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, 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.

It might be useful to have a figure here to illustrate 'branching' in a concrete example.

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

$$f(k^m) = mf(k). \tag{4.2}$$

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, Shannon considers another pair of integers ℓ and n such that

$$k^m \leq \ell^n \leq k^{m+1}, \tag{4.3}$$

or, taking logarithms,

$$\frac{m}{n} \leq \frac{\log \ell}{\log k} \leq \frac{m}{n} + \frac{1}{n}. \tag{4.4}$$

Now because the information measure $f(N)$ is monotonically increasing with N, the ordering in Eq. (4.3) means that

$$f(k^m) \leq f(\ell^n) \leq f(k^{m+1}), \tag{4.5}$$

specific situations. Again, we will come back to this.

and hence from Eq. (4.2) we obtain

$$mf(k) \leq nf(\ell) \leq (m+1)f(k). \tag{4.6}$$

Dividing through by $nf(k)$ we have

$$\frac{m}{n} \leq \frac{f(\ell)}{f(k)} \leq \frac{m}{n} + \frac{1}{n}, \tag{4.7}$$

which is very similar to Eq. (4.4). The trick is now that with k and ℓ fixed, we can choose an arbitrarily large value for n, so that $1/n = \epsilon$ is as small as we like. Then Eq. (4.4) is telling us that

$$\left|\frac{m}{n} - \frac{\log \ell}{\log k}\right| < \epsilon, \tag{4.8}$$

and hence Eq. (4.7) for the function $f(N)$ can similarly be rewritten as

$$\left|\frac{m}{n} - \frac{f(\ell)}{f(k)}\right| < \epsilon, \text{ or} \tag{4.9}$$

$$\left|\frac{f(\ell)}{f(k)} - \frac{\log \ell}{\log k}\right| \leq 2\epsilon, \tag{4.10}$$

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 is statistical mechanics: what are the units of entropy? In a physical chemistry course you might learn that entropy is measured in "entropy units," with the property that if you multiply by the absolute temperature (in Kelvin) you obtain an energy in units of calories per mole; this happens because the constant of proportionality is chosen to be the gas constant R, which refers to Avogadro's number of molecules.[3] In physics courses entropy is often defined with a factor of Boltzmann's constant k_B, so that if we multiply by the absolute temperature we again obtain an energy (in Joules) but now per molecule (or per degree of freedom), not per mole. In fact many statistical mechanics texts take the

[3] 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.

sensible view that temperature itself should be measured in energy units—that is, we should always talk about the quantity $k_B T$, 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. To make a bridge to this general problem from the simple case of equal probabilities, consider the situation where all the probabilities are rational, that is

$$p_{\rm n} = \frac{k_{\rm n}}{\sum_{\rm m} k_{\rm m}}, \tag{4.11}$$

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

Should have a figure to illustrate this grouping idea.

$$I_1 = I(\{p_{\rm n}\}). \tag{4.12}$$

At the second step, if we are in group n then we will gain $\log_2 k_{\rm n}$ 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

$$I_2 = \sum_{\rm n} p_{\rm n} \log_2 k_{\rm n}. \tag{4.13}$$

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

$$\begin{aligned} I_1 + I_2 &= \log_2\left(N_{\rm total}\right), && (4.14)\\ I(\{p_{\rm n}\}) + \sum_{\rm n} p_{\rm n}\log_2 k_{\rm n} &= \log_2\left(\sum_{\rm m} k_{\rm m}\right). && (4.15)\end{aligned}$$

Rearranging the terms, we find

$$\begin{aligned} I(\{p_{\rm n}\}) &= -\sum_{\rm n} p_{\rm n}\log_2\left(\frac{k_{\rm n}}{\sum_{\rm m} k_{\rm m}}\right) && (4.16)\\ &= -\sum_{\rm n} p_{\rm n}\log_2 p_{\rm n}. && (4.17)\end{aligned}$$

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 information obtained on hearing the answer to a question is measured uniquely by the entropy of the distribution of possible answers.

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

$$S_{\rm continuum} = -\int dx P(x)\log_2 P(x), \tag{4.18}$$

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

Problem 1: Dimensionality and the scaling of the entropy. As written, Eq (4.18) 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 $\rm n^{th}$ bin centered at $x_{\rm n}$. Then if the bins are reasonably small, the probability of falling in the $\rm n^{th}$ 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

$$S = -\sum_{\rm n} p_{\rm n} \log_2 p_{\rm n} = S_{\rm continuum} - \log_2(\Delta x). \tag{4.19}$$

More generally, show that in D dimensions

$$S = -\sum_{\rm n} p_{\rm n} \log_2 p_{\rm n} = S_{\rm continuum} - D\log_2(\Delta x). \tag{4.20}$$

The result in Eq (4.20) 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 in 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 Δ^2, 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 Δ. 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 0.05. Repeat the calculation of the entropy vs. box size. If you look closely enough (very small Δ) the underlying probability distribution really is two dimensional, since there is independent noise along x and y. But if your resolution is a little more coarse (large Δ) 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 $S(\Delta)$?

The problem of defining the entropy for continuous variables is familiar in statistical mechanics. In the simple example of an ideal gas in a finite box, we know that the quantum version of the problem has a discrete set of states, so that we can compute the entropy of the gas as a sum over these states. In the limit that the box is large, sums can be approximated as integrals, and if the temperature is high we expect that quantum effects are negligible and one might naively suppose that Planck's constant should disappear from the results; we recall that this is not quite the case. Planck's constant has units of momentum times position, and so is an elementary area for each pair of conjugate position and momentum variables in the classical phase space; in the classical limit the entropy becomes (roughly) the logarithm of the occupied volume in phase space, but this volume is measured in units of Planck's constant. If we had tried to start with a classical formulation (as did Boltzmann and Gibbs, of course) then we would find ourselves with the problems of Eq. (4.18), namely that we are trying to take the logarithm of a quantity with dimensions. If we measure phase space volumes in units of Planck's constant, then all is well.[4] 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 I need to do is write down this number. Note that the answers themselves can be very complex—different possible answers could correspond to lengthy essays, but the number of pages required to write these essays is irrelevant. If we agree in advance on the set of possible answers, all I have to do in answering the question is to provide a unique label. If we think of the label as a 'codeword' for the answer, then in this simple case the length of the codeword that represents the n^{th} possible answer is given by $\ell_{\mathrm{n}} = -\log_2 p_{\mathrm{n}}$, and the average length of a codeword is given by the entropy.

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

Clearly these remarks are interesting only if we can treat more general cases. Let us recall that in statistical mechanics we have the choice of working with a microcanonical ensemble, in which an ensemble of systems is distributed uniformly over states of fixed energy, or with a canonical ensemble, in which an ensemble of systems is distributed across states of different energies according to the Boltzmann distribution. The microcanonical ensemble

[4] In fact, when we start with quantum mechanics and pass to the classical limit, we pick up not just factors of Planck's constant to set the units of volume in phase space, but also a factor of $N!$, where N is the number of molecules. This is inherited from the quantum view in which particles are indistinguishable, and is important in resolving the mixing paradox, where the entropy of a gas seems to change if we just partition the container. Not crucial for the present discussion, but important more generally!

is like our simple example with all answers having equal probability: entropy really is just the log of the number of possible states. On the other hand, we know that in the thermodynamic limit there is not much difference between the two ensembles. This suggests that we can recover a simple notion of representing answers with codewords of length $\ell_{\rm n} = -\log_2 p_{\rm n}$ provided that we can find a suitbale analog of the thermodynamic limit.

Imagine that instead of asking a question once, we ask it many times. As an example, every day we can ask the weatherman for an estimate of the temperature at noon the next day. Now instead of trying to represent the answer to one question we can try to represent the whole stream of answers collected over a long period of time. Thus instead of a possible answer being labelled n, possible sequences of answers are labelled by $\rm n_1 n_2 \cdots n_{\it N}$. Of course these sequences have probabilites $P(\rm n_1 n_2 \cdots n_{\it N})$, and from these probabilities we can compute an entropy that must depend on the length of the sequence,

$$S(N) = -\sum_{\rm n_1}\sum_{\rm n_2}\cdots\sum_{{\rm n}_N} P({\rm n_1 n_2}\cdots {\rm n}_N)\log_2 P({\rm n_1 n_2}\cdots {\rm n}_N). \qquad (4.21)$$

Notice that we are *not* assuming that successive questions have independent answers, which would correspond to $P({\rm n_1 n_2}\cdots {\rm n}_N) = \prod_{i=1}^N p_{{\rm n}_i}$.

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

$$\lim_{N\to\infty}\frac{S(N)}{N} = \mathcal{S}, \qquad (4.22)$$

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

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

The idea of typical sequences, which is the information theoretic version of a thermodynamic limit, is enough to tell us that our simple arguments about representing answers by binary numbers ought to work on average for long sequences of answers. We will have to work significantly harder to show that this is really the smallest possible representation. An important if obvious consequence is that if we have many rather 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 an average of $\ell_{\rm n} = -\log_2 p_{\rm n}$ bits to represent each individual answer n. Thus answers which are more surprising require more space to write down.

[The previous paragraphs were a little vague, and helpful only if you have an intuition about statistical mechanics. At this point it would be good at least to give an example of how one can encode signals that have unequal probabilities, showing that we use codewords of different lengths. For now let me point you back to Shannon, Cover and Thomas, or even the charmingly idiosyncratic *Science and Information Theory* by Brillouin. More to come in a later revision.]