# Phys 446-546 Problem Set 3

Tue, 1 Mar 2005, covers lectures 8–10 (Entropy and Information Uncertainty), due 10 Mar (Note that this is a mix of lecture notes and problem sets in the sense that each problem has a long preamble.)

**Problem 1**: Recall that the entropy is defined as $S = k_B \ln \Omega$ where $\Omega$ is the number of states, and the Shannon information (or "information uncertainty") is $I = \log_2 \Omega = S/(k_B \ln 2)$. Consider a particle in a box of volume $V$ divided into $2^{12} = 4096$ distinguishable subvolumes in which the particle can be detected (i.e., a cube of volume $16^3$).

a) Suppose the particle is equally likely to be anywhere in the volume. What is the entropy $S$? How many bits of information are needed to characterize the particle location?

b) Now suppose the box is divided into 8 octants. What is the entropy and how many bits of info are needed to characterize the particle location if (i) the particle is known to be confined to just one of the octants, (ii) to two of the octants, or (iii) to four of the octants?

c) What is the probability $p_s$ that the particle is on the surface? (Hint: the volume of the interior is $14^3$.) Suppose that the particle really has only two physical states, surface or interior, with probabilities $p_s$ and $1 - p_s$. What then is the entropy and information uncertainty?

d) Suppose again that all $16^3$ individual volume elements are resolvable, but the particle has some greater probability of "sticking" to the surface, so there is only an overall probability of .1 that it will be found in the $14^3$ interior volume. What is the entropy in this situation? Suppose there is an ensemble of particles, half of which have "sticking" property, and the other half are surface-insensitive as in part a). What is the average entropy per particle and average information uncertainty per particle in this case?

**Problem 2**: The information uncertainty of an event in which there are $M$ equal probability possibilities (or states) $i = 1, \ldots, M$ is given by $I(M) = \log_2 M$. (This was determined by demanding that the information uncertainty for a situation with $M = M_1 \cdot M_2$ possibilities be given by $I(M_1 M_2) = I(M_1) + I(M_2)$, since the $M_1$ and $M_2$ possibilities could be considered as sequential, e.g., rolling a die [$M_1 = 6$] and flipping a coin [$M_2 = 2$], and so the uncertainty should be additive. The logarithm to the base 2 normalizes the result to binary bits.) This formula is easily generalized to the case of possibilities with unequal probabilities $p_i$ by mapping to a larger problem of $N$ events, in which each of the possible partitions of $N$ with $p_i N$ events in state $i$ is equal probability, so the earlier formula applies. The number of such partitions is $N!/\prod_{i=1}^{M} (p_i N)!$, so in the large $N$ limit the information uncertainty per original event becomes $I = \frac{1}{N} \log_2\big(\frac{N!}{\prod_i (p_i N)!}\big) \approx -\sum_{i=1}^{M} p_i \log_2 p_i$ (using

Stirling's formula, $N! \sim N \ln N - N$). For consistency, we note that this reduces to the earlier $I(M) = \log_2 M$ when each of the possibilities is equal probability, $p_i = 1/M$.*

The information gained in going from some initial probability distribution to some final distribution is the difference in the information uncertainties, $\Delta I = I_{\text{initial}} - I_{\text{final}}$.

a) Suppose as in the example given in class that one wakes up one morning knowing that there were three possibilities for the election results the previous day: $p_1 = 1/2$ that it is still unresolved, and $p_2 = p_3 = 1/4$ that either candidate A or B has been declared the winner.

(i) Suppose one sees only the latter part of the headline in the newspaper "... declared the winner", so that $p_1 = 0$ can be inferred, but the other two possibilities remain equal probability. How many bits of information have been obtained?

(ii) Suppose the newspaper is now fully uncovered so that it is known that candidate A has been declared the winner. How many bits of information have been gained in this step?

(iii) Suppose starting from the initial state the entire headline had been read in a single step – how many bits of information would be obtained in this way?

(iv) Suppose the initial probabilities had instead been $p_1 = p_2 = p_3 = 1/3$. How much information would be gained first by learning that $p_1 = 0$ and then that $p_2 = 1$?

b) Note that information is not always gained when possibilities are eliminated. Consider the initial case of $2^n + 1$ possibilities, with probabilities $p_0 = 1/2$ and $p_i = 1/2^{n+1}$ for $i = 1, \ldots, 2^n$ (note that $\sum_{i=0}^{2^n} p_i = 1$). Now suppose that case 0 is eliminated as a possibility, $p_0 = 0$, so that the remaining $2^n$ possibilities now have equal probability $1/2^n$. What is $\Delta I = I_{\text{initial}} - I_{\text{final}}$? For what values of $n$ is information gained or lost (information uncertainty decreases or increases, respectively) by eliminating 0 as a possibility?

* Another way to interpret the formula $I = -\sum_{i=1}^{M} p_i \log_2 p_i$ is to note that $\log_2 M = -\log_2(1/M) = -\log_2 p$ measures the information uncertainty associated to the case of equal probability possibilities, with $p = 1/M$. Hence $-\log_2 p_i$ can be regarded as measuring the uncertainty (or "surprise") associated to possibility $i$, with probability $p_i$. The formula for $I$ then just measures the average information uncertainty of the next symbol to be encountered, by summing $-\log_2 p_i$ over all possibilities, weighted by their associated probabilities $p_i$.

**Problem 3**: Genes in DNA consist of long strings of bases (A,C,G,T). When a gene is turned on, an RNA copy is made. (RNA has only a single strand, uses ribose instead of deoxyribose in the connecting backbone, and uses "U" instead of "T".) Ribosomes start the translation of the RNA into a protein at a pattern that has a cluster of G's and A's, then a gap, and then usually a start codon AUG, but sometimes GUG or UUG. The figure below shows ten examples of ribosome binding sites taken from E. Coli (running horizontally, with coordinates in the top row), showing the G/A cluster in the region around -10, and with the start codons aligned at position 0.

```
 -26  -20   -15  -10  -5    0   +5  +10  +15 +19
  |    |    |    |    |    |    |    |    |  |
 1 AAUUCCAUUCCUAGGAGGUUUGACCUGUGCGAGCUUUUAGUACUC
 2 UAGAGCCCUCAACCGGAGUUUGAAGCAUGGCUUCUAACUUUACUC
 3 UAAGGCAAUGCAAGGUCUCCUAAAAGAUGGAAACCCGAUUCCCUC
 4 CGGCCAUUCAAACAUGAGGAUUACCCAUGUCGAAGACAACAAAGA
 5 GUACUUCACUGAGUAUAAGAGGACAUUUGCCUAAAUUACCGCGUG
 6 CGUUGAAACUUUGGGUCAAUUUGAUCAUGGCAAAAUUAGAGACUG
 7 GUGCUUAGUAACUAAGGAUGAAAUGCAUGUCUAAGACAGCAUCUU
 8 GUUCAAGAUUGCUGGAGGCCUCCACUGUGAAAUCGCGUAGAGGCU
 9 GCUGCUAAAGGUCUAGGAGCUAAAGAAUGGAACAACUCACUAAAA
10 GAAGUUAACACUUUCGGAUAUUUCUGAUGAGUCGAAAAAUUAUCU
```

To quantify the information contained in these patterns, note that the "information uncertainty" per site in the absence of further information is $I = -\sum_{\alpha=\mathrm{A,C,G,T}} p_\alpha \log_2 p_\alpha = 4 \cdot (-\frac{1}{4}) \log_2 \frac{1}{4} = 2$ bits. Since position +1 above always has a U, the uncertainty at that site is reduced to zero, and hence the information content there is 2 bits (as expected for specifying a one of four choice). If a position has instead, say, only either A or G occurring with equal probability, then $I = -2 \cdot \frac{1}{2} \log_2 \frac{1}{2} = 1$ bit, and the information uncertainty decreases by 1 bit (as expected for specifying a two of four choice). In general, the information content at site $l$ can be regarded as the reduction in information uncertainty, in this case equal to $2 - I(l)$, where $I(l) = -\sum_{\alpha\in\{A,C,G,T\}} p_\alpha(l) \log_2 p_\alpha(l)$ can be any real number between 0 and 2 bits.*

a) What is the information content at sites 0,1,2 in the above figure?

b) What is the information content at sites -11,-10,-9 in the above figure?

c) Find a site with extremely low (or lowest if possible) information content

* From the standpoint of the ribosome, before binding each of its "fingers" sees 4 indistinguishable possibilities and is hence "uncertain" by 2 bits. After binding, the uncertainty at each finger is lower. If only 1 base ever binds, then the final uncertainty is 0 bits. The decrease in uncertainty of 2 bits is thus a measure of the sequence conservation or information at the binding site. If a finger accepts 2 bases with equal probability, then the uncertainty remaining is 1 bit, and the information is 1 bit. When a "finger" accepts all 4 bases with equal probability, it's not really doing anything and requires 0 bits of information in sequence conservation.

(Note that the total information content in the regions depicted in the figure, i.e., $I(l)$ summed over sites, has to be at least greater than the number of bits required to specify the location of the binding sites within the E.Coli genome — otherwise the ribosome would never be able to find the binding sites in a dynamic process.)

**Problem 4**: Suppose a source emits a stream of binary digits (0's and 1's) with probabilities $p_0 = x$ and $p_1 = 1 - x$.

(a) What is the information uncertainty as function of $x$?

(b) What is average number of bits per above binary digit for each of the four values $x = 1/2,\ 3/4,\ \sqrt{2}/2,\ 9/10$. Recall that this also specifies the minimum number of bits required to transmit the same information content.

(c) The result of (b) suggests that some compression scheme should be possible for some values of $x$. Consider a simple scheme in which we encode two binary digits at a time into a new binary symbol taking values $\alpha, \beta$. For example, if 0 is most likely, we could take $00 = \alpha$, $01 = \beta\alpha$, $10 = \beta\beta\alpha$, and $11 = \beta\beta\beta$. (Note that it is possible to decode the stream of $\alpha, \beta$'s uniquely back to the original stream of 0,1's.) (i) What is the probability of $\alpha$ and $\beta$ in the new scheme as a function of $x$, and what is the information uncertainty per symbol in the new scheme? (ii) For each of the four values of $x$ in (b), how does the information per symbol in the new scheme compare to the theoretical minimum? (iii) How could this scheme be modified to come even closer to the theoretical minimum?

**Problem 5**: Recall that the genomic code uses a sequence of three bases (each which can take the four values A,C,G,T) to code for each of the 20 amino acids that are assembled to make proteins.

(i) Suppose that there were some advantage for an organism to have a compressed genome and to code for these proteins with the minimum number of bases. If the probabilities of the amino acids in such an organism are $p_1 = p_2 = p_3 = p_4 = 1/8$ and $p_5 = \ldots = p_{20} = 1/32$, what is the theoretical minimum average number of bases (A,C,G,T) that could be used to code for these proteins. Find a scheme that comes close to this theoretical minimum.

(ii) Repeat for the case $p_1 = p_2 = 1/4$, $p_3 = p_4 = p_5 = p_6 = 1/16$, $p_7 = \ldots = p_{20} = 1/(4 \cdot 14)$.