Physics 446-546, Problem Set 5
(issued Nov 18, '03, due 2 Dec '03)

#1 (from Bialek notes, physics/0205030, p. 32)

Problem 6: 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, $P(T = T_*) << 1$. 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 [Ginzburg I., & Sejnowski, T. J. (1996). Dynamics of Rule Induction by Making Queries: Transition Between Strategies, in *18th Annual Conference of the Cognitive Science Society,* pp. 121–125 (Lawrence Erlbaum, Mahwah NJ). See also http://www.cnl.salk.edu/CNL/annual–reps/annual–rep95.html]. Related but distinct criteria for optimal experimental design have been developed in the statistical literature [Fedorov, V. V. (1972). *Theory of Optimal Experimental Design,* translated and edited by Studden, W. J., & Klimko, E. M. (Academic Press, New York)].)

#2 (from Bialek notes, physics/0205030, p. 33)

Problem 7: Positivity of information.

$$I(D \to W) = S(W) - \sum_D P(D) S(W|D) \tag{43}$$

$$= \sum_W \sum_D P(W, D) \log_2 \left[\frac{P(W, D)}{P(W) P(D)} \right]. \tag{44}$$

Prove the above formula and prove that the mutual information $I(D \to W)$, defined in Eq. (44), is positive.

#3 (from Bialek notes, physics/0205030, p. 35)

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

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

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

$$\langle f_0\rangle = \sum_D P(D) = 1.$$

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

The problem of maximizing a quantity subject to constraints is formulated using Lagrange multipliers. The result is that

$$P(D) = \frac{1}{Z} \exp\left[-\sum_{\rm i=1}^{K} \lambda_{\rm i} f_{\rm i}(D)\right], \tag{53}$$

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

Problem 8: Details. Derive Eq. (53). In particular, show that Eq. (53) provides a probability distribution which genuinely *maximizes* the entropy, rather than being just an extremum.

(Recall "Your Turn 6G" on p.224 of the course text, and 6.1′ on p.232, for a similar derivation of the Boltzmann distribution.)

#4

(a) In class, we discussed the "occasionally dishonest casino" that used two kinds of dice: 99% were fair, but 1% were loaded so that a six came up 50% of the time. The conditional probabilities were thus P(six|$D_{\rm loaded}$)= 1/2, P(six|$D_{\rm fair}$) = 1/6. If we then pick a die at random, what are the joint probabilities P(six,$D_{\rm loaded}$) and P(six,$D_{\rm fair}$)? What is the probability of rolling a six from the die we picked up? If we rolled three sixes in a row, we saw that the posterior probability P($D_{\rm loaded}$|3 sixes) that it was loaded was only 3/14. How many sixes in a row would we have to roll before concluding it was more likely to have been a loaded die?

(b) In class we also discussed the case of "the rare genetic disease" carried by only one in a million people, and a screening test that is 100% sensitive (always correct if you have the disease) and 99.99% specific (gives a false positive only 0.01% of the time), and concluded it would not be sensible to take the test. For that problem, we thus had P(D) = 10^{-6}, P(+|D) = 1, P(+|$\overline{\rm D}$) = 10^{-4}, where D,$\overline{\rm D}$ denote diseased and not diseased, resp., and + indicates positive on the screening test. For general values of P(D), P(+|D), P(+|$\overline{\rm D}$), determine the conditional probability P(D|+) of being diseased if the test result is positive, and characterize when it is greater than the risk P(D) of having the disease.

#5

In class we discussed the ferromagnetic 2D Ising model on a square lattice with partition function $Z(K) = \sum_{\{\sigma_i=\pm1\}} \exp(\sum_{<ij>} K\sigma_i\sigma_j)$, where the $<ij>$ sum is over nearest neighbor pairs, and showed that $Z(K) \propto Z(L)$ where $(e^{2L}-1)(e^{2K}-1) = 2$. Generalize the argument given in class to the case of an anisotropic model, with couplings K_h on the horizontal links and K_v on the vertical links . Discuss the phase diagram in the K_h, K_v plane and locate the location of the transition based on the extended duality relation. (Hint: it is now a line, rather than a single point.)

#6

In class, we discussed the Hopfield model for associative memory.

(a) Consider first the case of a neural net with $N = 8$ nodes and a single pattern $\xi_i^{(1)} = 1$ for all i, i.e., the vector $\vec{\xi}^{(1)} = (1,1,1,1,1,1,1,1)$. Write down the matrix $w_{ij} = \frac{1}{N}\xi_i^{(1)}\xi_j^{(1)}$ and determine its basins of attraction for test vectors $\vec{\zeta}$ in terms of the coordinates of ζ.

(b) Now add a second pattern $\vec{\xi}^{(2)} = (1,1,1,1,-1,-1,-1,-1)$, write down $w_{ij} = \frac{1}{N}\sum_{\mu=1}^{p}\xi_i^{(\mu)}\xi_j^{(\mu)}$ for $p = 2$ and determine the basins of attraction for test vectors $\vec{\zeta}$.

(c) Add a third pattern $\vec{\xi}^{(3)} = (-1,1,1,1,1,1,1,1,-1)$, again write down $w_{ij} = \frac{1}{N}\sum_{\mu=1}^{p}\xi_i^{(\mu)}\xi_j^{(\mu)}$ now for $p = 3$, and determine its basins of attraction for test vectors $\vec{\zeta}$.