Appendix A: Appendix

In these sections I collect things which are off the to side, or in the background, of the main arguments made in the text. Some readers will find the background essential, others will find the asides more interesting than what I thought were the main points. I hope, however, that everyone finds something useful here. As with the main text, I try not to skip steps, and problems are embedded in the narrative. To a large extent, the Appendices are unedited as of September 18, 2011; for some of the newer ones, especially, much work is needed.

1. Poisson processes

Photons from a conventional light source arrive at a detector as a random process, specifically a Poisson process. The defining feature of the Poisson process is that each event (photon arrival) is independent of all the others, given that we know the rate $r(t)$ at which the events occur. In these notes we'll go through the detailed consequences of this simple assumption of independence; hopefully some of the results are familiar. Note that many textbook presentations make a big deal out of the distinction between a "homogeneous" Poission process, in which the rate is a constant, $r(t) = \bar{r}$, and an "inhomogeneous" Poisson process in which it can depend on time. The general case isn't that hard, so I prefer to start there.

One should perhaps note at the outset that most light sources are not exactly Poisson, but the approximation is very good. There are many more systems for which the Poisson model is a decent if not excellent approximation, and so we'll discuss all this without further reference to photons: we are describing the statistics of arbitrary point events which occur at times $t_1, t_2, \cdots, t_N$.

The rate $r(t)$ can be thought of either as the mean rate of events that we would observe in the neighborhood of time t if we did the same experiment many times, or equivalently as the probability per unit time that we observe an event at t. Recall that there is the same dual definition for the concentration $c(\mathbf{x})$ of moelcules—either the mean number of molecule per unit volume that we find in the neighborhood of a point $\mathbf{x}$, or the probability per unit volume that we observe a single molecule at $\mathbf{x}$.

Since the events are independent, the probability density for observing events at times $t_1, t_2, \cdots, t_N$ must be proportional to a product of the rates evaluated at these times,

$$P[\{t_i\}|r(\tau)] \propto r(t_1)r(t_2)\cdots r(t_N) \equiv \prod_{\mathrm{i}=1}^{N} r(t_\mathrm{i}). \tag{A1}$$

But to get the exact form of the distribution we must include a factor that measures the probability of *no* events occurring at any other times. The probability of an event occurring in a small bin of size $\Delta\tau$ surrounding time t is, by the original definition of the rate, $p(t) = r(t)\Delta\tau$, so the probability of no event must be $1 - p(t)$. Thus we need to form a product of factors $1-p(t)$ for all times not equal to the special t_i where we observed events. Let's call this factor F,

$$F = \prod_{\mathrm{n}\neq\mathrm{i}} [1 - p(t_\mathrm{n})]. \tag{A2}$$

Then the probability of observing events in bins surrounding the t_i is

$$P[\{t_\mathrm{i}\}|r(\tau)](\Delta\tau)^N = \frac{1}{N!} F \prod_{\mathrm{i}=1}^{N} [r(t_\mathrm{i})\Delta\tau], \tag{A3}$$

where the $N!$ corrects for all the different ways of assigning labels $1, 2, \cdots, N$ to the events we observe.

To proceed we pull out all the factors related to the t_i and isolate the terms independent of these times:

$$\begin{aligned} P[\{t_i\}|r(\tau)](\Delta\tau)^N &= \frac{1}{N!} F \prod_{\mathrm{i}=1}^{N} [r(t_\mathrm{i})\Delta\tau] \\ &= \frac{1}{N!} \prod_{\mathrm{n}\neq\mathrm{i}} [1 - r(t_\mathrm{n})\Delta\tau] \prod_{\mathrm{i}=1}^{N} [r(t_\mathrm{i})\Delta\tau] \qquad \text{(A4)} \\ &= \frac{1}{N!} \prod_{\mathrm{n}} [1 - r(t_\mathrm{n})\Delta\tau] \prod_{\mathrm{i}=1}^{N} \left[\frac{r(t_\mathrm{i})\Delta\tau}{1 - r(t_\mathrm{i})\Delta\tau}\right]; \qquad \text{(A5)} \end{aligned}$$

keep in mind that $\prod_\mathrm{n}$ denotes a product over *all* possible times t_n.

To simplify Eq (A5) we remember that products can be turned into sums by taking logarithms, so that

$$\prod_\mathrm{n} [1 - r(t_\mathrm{n})\Delta\tau] = \exp\left(\sum_\mathrm{n} \ln [1 - r(t_\mathrm{n})\Delta\tau]\right). \tag{A6}$$

Now when we substitute back into Eq (A5) we find

$$\begin{aligned}P[\{t_{\rm i}\}|r(\tau)](\Delta\tau)^N &= \frac{1}{N!}\prod_{\rm n}\left[1-r(t_{\rm n})\Delta\tau\right]\prod_{\rm i=1}^{N}\left[\frac{r(t_{\rm i})\Delta\tau}{1-r(t_{\rm i})\Delta\tau}\right] \\ &= \frac{1}{N!}\exp\left(\sum_{\rm n}\ln\left[1-r(t_{\rm n})\Delta\tau\right]\right)\prod_{\rm i=1}^{N}\left[\frac{r(t_{\rm i})\Delta\tau}{1-r(t_{\rm i})\Delta\tau}\right].\end{aligned} \tag{A7}$$

We are interested in the case where the time bin $\Delta\tau$ is very small (we introduced these artificially, remember), which means that we need to take the logarithm of numbers that are almost equal to one. We recall that the Taylor series of the logarithm is

$$\ln(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \cdots . \tag{A8}$$

In this case we apply this expansion to

$$\ln\left[1-r(t_{\rm n})\Delta\tau\right] = -r(t_{\rm n})\Delta\tau - \frac{1}{2}\left[r(t_{\rm n})\Delta\tau\right]^2 + \cdots , \tag{A9}$$

so our expression for the probability can be written as

$$\begin{aligned}P[\{t_{\rm i}\}|r(\tau)](\Delta\tau)^N &= \frac{1}{N!}\exp\left(\sum_{\rm n}\ln\left[1-r(t_{\rm n})\Delta\tau\right]\right)\prod_{\rm i=1}^{N}\left[\frac{r(t_{\rm i})\Delta\tau}{1-r(t_{\rm i})\Delta\tau}\right] \\ &= \frac{1}{N!}\exp\left(\sum_{\rm n}[-r(t_{\rm n})\Delta\tau] - \frac{1}{2}\sum_{\rm n}[-r(t_{\rm n})\Delta\tau]^2 + \cdots\right)\prod_{\rm i=1}^{N}\left[\frac{r(t_{\rm i})\Delta\tau}{1-r(t_{\rm i})\Delta\tau}\right].\end{aligned} \tag{A10}$$

This expression involves a sum over bins, with factors of the bin width $\Delta\tau$. We recall that this converges, as the bins become small, to an integral:

$$\lim_{\Delta\tau\to 0}\sum_{\rm n} f(t_{\rm n})\Delta\tau = \int dt\, f(t), \tag{A11}$$

for any smooth function $f(t)$. In the present case this means that

$$\lim_{\Delta\tau\to 0}\exp\left(\sum_{\rm n}[-r(t_{\rm n})\Delta\tau] - \frac{1}{2}\sum_{\rm n}[-r(t_{\rm n})\Delta\tau]^2 + \cdots\right) = \exp\left[-\int dt\, r(t) - \frac{1}{2}\Delta\tau\int dt\, r^2(t) + \cdots\right]. \tag{A12}$$

Now we notice that the second integral in the exponential has an extra factor of $\Delta\tau$, which comes from the $(\Delta\tau)^2$ in the previous expression, but if we really let $\Delta\tau$ go to zero this must be negligible as long as the rate doesn't become infinite.

Similarly, we have in Eq (A10) factors like

$$\frac{r(t_{\rm i})\Delta\tau}{1-r(t_{\rm i})\Delta\tau},$$

and again as $\Delta\tau\to 0$ we can expand this in powers of $\Delta\tau$ and drop all but the first term. This is equivalent to replacing the denominator of the fraction by 1. So, when the dust clears, the expression for the probability density of the event times becomes

$$P[\{t_i\}|r(\tau)] = \frac{1}{N!}\exp\left[-\int_0^T dt\; r(t)\right]\prod_{\rm i=1}^{N} r(t_{\rm i}), \tag{A13}$$

where we have set the limits on the integral to refer to the whole duration of our observations, from $t=0$ to $t=T$. Note that this is a probability density for the N arrival times t_1, t_2, $\cdots$, t_N and hence has units $(\text{time})^{-N}$.

It is a useful exercise to check the normalization of the probability distribution in Eq. (A13). We want to calculate the total probability, which involves taking the term with N events and integrating over all N arrival times, then summing on N. Let's call this sum Z,

$$Z \equiv \sum_{N=0}^{\infty} \int_0^T dt_1 \int_0^T dt_2 \cdots \int_0^T dt_N P[\{t_{\rm i}\}|r(t)] \tag{A14}$$

$$= \sum_{N=0}^{\infty} \int_0^T dt_1 \int_0^T dt_2 \cdots \int_0^T dt_N \frac{1}{N!} \exp\left[-\int_0^T dt\, r(t)\right] \prod_{\rm i=1}^{N} r(t_{\rm i}). \tag{A15}$$

Notice that the exponential does not depend on the $\{t_{\rm i}\}$ or on N, so we can take it outside the sum and integral. Furthermore, although we have to integrate over all the N different $t_{\rm i}$ together (an N dimensional integral), the integrand is just a product of terms that depend on each individual $t_{\rm i}$. This means that really we have a product of N one dimensional integrals:

$$Z = \exp\left[-\int_0^T dt\, r(t)\right] \sum_{N=0}^{\infty} \frac{1}{N!} \int_0^T dt_1 \cdots \int_0^T dt_N\, r(t_1) \cdots r(t_N) \tag{A16}$$

$$= \exp\left[-\int_0^T dt\, r(t)\right] \sum_{N=0}^{\infty} \frac{1}{N!} \int_0^T dt_1\, r(t_1) \int_0^T dt_2\, r(t_2) \cdots \int_0^T dt_N\, r(t_N) \tag{A17}$$

$$= \exp\left[-\int_0^T dt\, r(t)\right] \sum_{N=0}^{\infty} \frac{1}{N!} \left[\int_0^T dt\, r(t)\right]^N . \tag{A18}$$

Recall that the series expansion of the exponential function is

$$\exp(x) = \sum_{N=0}^{\infty} \frac{1}{N!} x^N , \tag{A19}$$

so we can actually do the sum in Eq. (A18):

$$\exp\left[-\int_0^T dt\, r(t)\right] \sum_{N=0}^{\infty} \frac{1}{N!} \left[\int_0^T dt\, r(t)\right]^N = \exp\left[-\int_0^T dt\, r(t)\right] \times \exp\left[+\int_0^T dt\, r(t)\right] \tag{A20}$$

$$= 1, \tag{A21}$$

which completes our check on the normalization of the distribution.

Next we would like to derive an expression for the distribution of counts, which we write as $P(N|\langle N\rangle)$ to remind us that the shape of the distribution depends (as we will see) only on its mean. To do this we take the full probability distribution $P[\{t_i\}|r(\tau)]$, pick out the term involving N events, and then integrate over all the possible arrival times of these events:

$$P(N|\langle N\rangle) = \int_0^T dt_1 \cdots \int_0^T dt_N P[\{t_{\rm i}\}|r(\tau)] \tag{A22}$$

$$= \int_0^T dt_1 \cdots \int_0^T dt_N \frac{1}{N!} \exp\left[-\int_0^T dt\, r(t)\right] \prod_{i=1}^{N} r(t_{\rm i}). \tag{A23}$$

As in the discussion leading to Eq. (A18) we notice that the exponential factor can be taken outside the integral, and that we have a product of N one dimensional integrals rather than a full N dimensional integral:

$$\begin{aligned} P(N|\langle N\rangle) &= \int_0^T dt_1 \cdots \int_0^T dt_N \frac{1}{N!} \exp\left[-\int_0^T dt\, r(t)\right] \prod_{\mathrm{i}=1}^{N} r(t_\mathrm{i}) \\ &= \frac{1}{N!} \exp\left[-\int_0^T dt\, r(t)\right] \int_0^T dt_1 \cdots \int_0^T dt_N \prod_{i=1}^{N} r(t_i) \\ &= \frac{1}{N!} \exp\left[-\int_0^T dt\, r(t)\right] \left[\int_0^T dt\, r(t)\right]^N \qquad \text{(A24)} \\ &\equiv \frac{1}{N!} \exp(-Q) Q^N, \qquad \text{(A25)} \end{aligned}$$

where we have defined

$$Q = \int_0^T dt\, r(t). \qquad \text{(A26)}$$

In particular, the probability that no events occur in the time from $t = 0$ to $t = T$ is $P(0) = \exp(-Q)$, or

$$P(0|\langle N\rangle) = \exp\left[-\int_0^T dt\, r(t)\right]. \qquad \text{(A27)}$$

With the probability distribution of counts from Eq. (A25), we can compute the mean and the variance of the count. To obtain the mean we compute

$$\begin{aligned} \langle N\rangle &\equiv \sum_{N=0}^{\infty} P(N) N \qquad \text{(A28)} \\ &= \sum_{N=0}^{\infty} \frac{1}{N!} \exp(-Q) Q^N N \qquad \text{(A29)} \\ &= \exp(-Q) \sum_{N=0}^{\infty} \frac{1}{N!} Q^N N. \qquad \text{(A30)} \end{aligned}$$

Now we have already made use of the series expansion for the exponential, Eq. (A19), and to sum this last series we notice that

$$Q^N N = Q \frac{\partial}{\partial Q} Q^N, \qquad \text{(A31)}$$

so that

$$\begin{aligned} \langle N\rangle &= \exp(-Q) \sum_{N=0}^{\infty} \frac{1}{N!} Q^N N \\ &= \exp(-Q) \sum_{N=0}^{\infty} \frac{1}{N!} Q \frac{\partial}{\partial Q} Q^N \qquad \text{(A32)} \\ &= \exp(-Q) Q \frac{\partial}{\partial Q} \sum_{N=0}^{\infty} \frac{1}{N!} Q^N \qquad \text{(A33)} \\ &= \exp(-Q) Q \frac{\partial}{\partial Q} \exp(+Q), \qquad \text{(A34)} \end{aligned}$$

where in the last step we recognize the series for the exponential. Now the derivative of the exponential is just the exponential itself,

$$\frac{\partial}{\partial Q} \exp(+Q) = \exp(+Q), \qquad \text{(A35)}$$

so that

$$\begin{aligned} \langle N\rangle &= \exp(-Q) Q \frac{\partial}{\partial Q} \exp(+Q) \\ &= \exp(-Q) Q \exp(+Q) = Q. \qquad \text{(A36)} \end{aligned}$$

We see that the mean count is what we have called Q, the integral of the rate.

Now we can write the count distribution directly in terms of its mean:

$$P(N|\langle N\rangle) = \exp(-\langle N\rangle) \frac{\langle N\rangle^N}{N!}, \qquad \text{(A37)}$$

which is what we need to start the discussion of photon counting in vision, Eq (**??**).

We can do a very similar calculation to find the variance of the count distribution. We start by computing the average of N^2,

$$\langle N^2\rangle = \sum_{N=0}^{\infty} N^2 P(N). \qquad \text{(A38)}$$

Substituting for $P(N)$ from Eq. (A25) and rearranging, we have

$$\begin{aligned} \langle N^2\rangle &= \sum_{N=0}^{\infty} N^2 P(N) \\ &= \sum_{N=0}^{\infty} N^2 \exp(-Q) \frac{1}{N!} Q^N \qquad \text{(A39)} \\ &= \exp(-Q) \sum_{N=0}^{\infty} \frac{1}{N!} N^2 Q^N. \qquad \text{(A40)} \end{aligned}$$

The trick is once again to write the extra factors of N (here N^2) in terms of derivatives with respect to Q. Now we know that

$$\frac{\partial^2}{\partial Q^2} Q^N = N(N-1) Q^{N-2}, \qquad \text{(A41)}$$

so we can write

$$Q^2 \frac{\partial^2}{\partial Q^2} Q^N = (N^2 - N) Q^N, \qquad \text{(A42)}$$

which is almost what we want. But we can use the formula in Eq. (A31) to finish the job, obtaining

$$N^2 Q^N = Q^2 \frac{\partial^2}{\partial Q^2} Q^N + Q \frac{\partial}{\partial Q} Q^N. \qquad \text{(A43)}$$

Now we can substitute into Eq. (A40) and follow the steps corresponding to Eq's (A32) through (A36):

$$\begin{aligned}
\langle N^2 \rangle &= \exp(-Q) \sum_{N=0}^{\infty} \frac{1}{N!} N^2 Q^N \\
&= \exp(-Q) \sum_{N=0}^{\infty} \frac{1}{N!} \left[Q^2 \frac{\partial^2}{\partial Q^2} Q^N + Q \frac{\partial}{\partial Q} Q^N \right] && \text{(A44)} \\
&= \exp(-Q) Q^2 \frac{\partial^2}{\partial Q^2} \sum_{N=0}^{\infty} \frac{1}{N!} Q^N + \exp(-Q) Q \frac{\partial}{\partial Q} \sum_{N=0}^{\infty} \frac{1}{N!} Q^N && \text{(A45)} \\
&= \exp(-Q) Q^2 \frac{\partial^2}{\partial Q^2} \exp(+Q) + \exp(-Q) Q \frac{\partial}{\partial Q} \exp(+Q) && \text{(A46)} \\
&= \exp(-Q) Q^2 \exp(+Q) + \exp(-Q) Q \exp(+Q) && \text{(A47)} \\
&= Q^2 + Q. && \text{(A48)}
\end{aligned}$$

Now since we have already identified Q as equal to the mean count, this means that the mean square count can be written as

$$\langle N^2 \rangle = \langle N \rangle^2 + \langle N \rangle. \qquad \text{(A49)}$$

But the variance of the count is defined by

$$\begin{aligned}
\langle (\delta N)^2 \rangle &\equiv \langle N^2 \rangle - \langle N \rangle^2 && \text{(A50)} \\
&= [\langle N \rangle^2 + \langle N \rangle] - \langle N \rangle^2 = \langle N \rangle. && \text{(A51)}
\end{aligned}$$

Thus the variance of the count for a Poisson process is equal to the mean count.

The next characteristic of the Poisson process is the interval between events. The probability per unit time that we observe an event at time t is given by the rate, $r(t)$. The probability that we observe no events in the interval $[t, t+\tau)$ is given by

$$P(0) = \exp\left[-\int_t^{t+\tau} dt'\, r(t') \right]. \qquad \text{(A52)}$$

The probability per unit time that this interval is closed by an event is again the rate, now at time $t+\tau$. Thus the probability per unit time that we see events at t and $t+\tau$, with no events in between is given by

$$P(t, t+\tau) = r(t) \exp\left[-\int_t^{t+\tau} dt'\, r(t') \right] r(t+\tau). \qquad \text{(A53)}$$

In the simple case that the rate is constant, this is just $P(t, t+\tau) = r^2 e^{-r\tau}$. On the other hand, if the rate varies, the average probability for observing two events separated by an empty interval of duration τ is

$$P_2(\tau) = \left\langle r(t) \exp\left[-\int_t^{t+\tau} dt'\, r(t') \right] r(t+\tau) \right\rangle, \qquad \text{(A54)}$$

where $\langle \cdots \rangle$ is an average over these variations in rate.

If we ask for the probability density of intervals, this is really the conditional probability that the next event will be at $t+\tau$ given that there was an event at t. To form this conditional probability we need to divide by the probability of an event at t, but this is just the average rate. Again, in the simple case of constant rate, this yields the probability density of inter–event intervals,

$$p(\tau) = r e^{-r\tau}. \qquad \text{(A55)}$$

This exponential form is one of the classic signatures of a Poission process. We can think of it as arising because the moment at which the interval closes has no memory of the moment at which it opened, and so the probability that there has not ben an event must be a product of terms for the absence of an event in each small time slice $\Delta\tau$, as in the derivation above, and this product becomes an exponential.

Our last task is to evaluate averages over Poisson processes, such as the one in Eq (33),

$$\left\langle \sum_{\rm i} V_0(t-t_{\rm i}) \right\rangle = \sum_{N=0}^{\infty} \int_0^T dt_1 \cdots \int_0^T dt_N P[\{t_{\rm i}\}|r(t)] \sum_{\rm i} V_0(t-t_{\rm i}). \tag{A56}$$

We proceed simply and systematically, looking at one term in our sum and doing the integrals one at a time.

One term in the sum means that we choose, for example i = 1 *and* one particular value of N. This term is

$$\int_0^T dt_1 \cdots \int_0^T dt_N P[\{t_{\rm i}\}|r(t)] V_0(t-t_1) = \int_0^T dt_1 \cdots \int_0^T dt_N \exp\left[-\int_0^T d\tau\, r(\tau)\right] \frac{1}{N!} r(t_1) r(t_2) \cdots r(t_N) V_0(t-t_1). \tag{A57}$$

Notice that the exponential factor (along the the $1/N!$) is constant and comes outside the integral. Now we rearrange the order of the integrals:

$$\int_0^T dt_1 \int_0^T dt_2 \cdots \int_0^T dt_N r(t_1) r(t_2) \cdots r(t_N) V_0(t-t_1) = \int_0^T dt_1\, r(t_1) V_0(t-t_1) \int_0^T dt_2\, r(t_2) \cdots \int_0^T dt_N\, r(t_N) \tag{A58}$$

$$= \left[\int_0^T dt_1\, r(t_1) V_0(t-t_1)\right] \left[\int_0^T d\tau r(\tau)\right]^{N-1}. \tag{A59}$$

But the fact that we chose i = 1 was arbitrary; we would have gotten the same answer for any $\rm i = 1, 2, \cdots, N$. Thus summing over i is the same as multiplying by N. This leaves us with the sum on N, so we put everything back together to find

$$\left\langle \sum_{\rm i} V_0(t-t_{\rm i}) \right\rangle = \exp\left[-\int_0^T d\tau\, r(\tau)\right] \left[\int_0^T dt_1\, r(t_1) V_0(t-t_1)\right] \sum_{N=0}^{\infty} \frac{N}{N!} \left[\int_0^T d\tau\, r(\tau)\right]^{N-1} \tag{A60}$$

$$= \exp\left[-\int_0^T d\tau\, r(\tau)\right] \int_0^T dt_1\, r(t_1) V_0(t-t_1) \sum_{N=0}^{\infty} \frac{1}{N!} \left[\int_0^T d\tau\, r(\tau)\right]^{N} \tag{A61}$$

$$= \exp\left[-\int_0^T d\tau\, r(\tau)\right] \int_0^T dt_1\, r(t_1) V_0(t-t_1) \exp\left[+\int_0^T d\tau\, r(\tau)\right] \tag{A62}$$

$$= \int_0^T dt_1\, r(t_1) V_0(t-t_1). \tag{A63}$$

Thus what we have shown is that our simple model of summing pulses from single photons generates a voltage that responds linearly to the light intensity,

$$\langle V(t) \rangle = V_{\rm DC} + \int dt' V_0(t-t') r(t'), \tag{A64}$$

which is Eq (34) in the main text.

Actually, we have shown something more general, which will be useful below. The expectation value we have computed is of the form

$$\left\langle \sum_{\rm i} f(t_{\rm i}) \right\rangle. \tag{A65}$$

What we have seen is that summing over arrival times is, on average, equivalent to integrating over the rate,

$$\left\langle \sum_{\rm i} f(t_{\rm i}) \right\rangle = \int d\tau\, r(\tau) f(\tau). \tag{A66}$$

Intuitively, this makes sense: the sum over arrival times approximates a density along the time axis, and this density is the rate, with units of (events)/(time).

Now we need to do the same calculation, but for the correlation function of the voltage. Again we have

$$V(t) = \sum_{\rm i} V_0(t-t_{\rm i}), \tag{A67}$$

and we want to compute $\langle V(t) V(t') \rangle$. Intuitively, the arrival times of photons are independent of one another—

this is the essence of the Poisson process—and so we should have

$$\langle V(t)V(t')\rangle = \left\langle \sum_{\rm i} V_0(t-t_{\rm i})\sum_{\rm j} V_0(t'-t_{\rm j})\right\rangle \tag{A68}$$

$$= \sum_{\rm ij}\langle V_0(t-t_{\rm i})V_0(t'-t_{\rm j})\rangle \tag{A69}$$

$$= \sum_{\rm i\neq j}\langle V_0(t-t_{\rm i})V_0(t'-t_{\rm j})\rangle + \sum_{\rm i}\langle V_0(t-t_{\rm i})V_0(t'-t_{\rm i})\rangle \tag{A70}$$

$$= \sum_{\rm i\neq j}\langle V_0(t-t_{\rm i})\rangle\langle V_0(t'-t_{\rm j})\rangle + \sum_{\rm i}\langle V_0(t-t_{\rm i})V_0(t'-t_{\rm i})\rangle, \tag{A71}$$

where we use the independence of $t_{\rm i}$ and $t_{\rm j}$ for $\rm i \neq j$ in the last step. It's useful to add and subtract the "diagonal" $\rm i = j$ term from the sum, so that

$$\langle V(t)V(t')\rangle = \sum_{\rm i\neq j}\langle V_0(t-t_{\rm i})\rangle\langle V_0(t'-t_{\rm j})\rangle + \sum_{\rm i}\langle V_0(t-t_{\rm i})\rangle\langle V_0(t'-t_{\rm i})\rangle$$

$$+ \sum_{\rm i}\langle V_0(t-t_{\rm i})V_0(t'-t_{\rm i})\rangle - \sum_{\rm i}\langle V_0(t-t_{\rm i})\rangle\langle V_0(t'-t_{\rm i})\rangle \tag{A72}$$

$$= \sum_{\rm ij}\langle V_0(t-t_{\rm i})\rangle\langle V_0(t'-t_{\rm j})\rangle + \sum_{\rm i}\left[\langle V_0(t-t_{\rm i})V_0(t'-t_{\rm i})\rangle - \langle V_0(t-t_{\rm i})\rangle\langle V_0(t'-t_{\rm i})\rangle\right]. \tag{A73}$$

The key step now is to notice that we can rearrange the sums and expectation values in the first term,

$$\sum_{\rm ij}\langle V_0(t-t_{\rm i})\rangle\langle V_0(t'-t_{\rm j})\rangle = \sum_{\rm i}\langle V_0(t-t_{\rm i})\rangle\sum_{\rm j}\langle V_0(t'-t_{\rm j})\rangle \tag{A74}$$

$$= \left\langle \sum_{\rm i} V_0(t-t_{\rm i})\right\rangle\left\langle \sum_{\rm j} V_0(t'-t_{\rm j})\right\rangle \tag{A75}$$

$$= \langle V(t)\rangle\langle V(t')\rangle, \tag{A76}$$

where in the last step we recognize the voltage itself, from Eq (A67). Thus Eq (A73) can be rewritten as an equation for the covariance of the voltage fluctuations,

$$\langle \delta V(t)\delta V(t')\rangle \equiv \langle V(t)V(t')\rangle - \langle V(t)\rangle\langle V(t')\rangle = \sum_{\rm i}\left[\langle V_0(t-t_{\rm i})V_0(t'-t_{\rm i})\rangle - \langle V_0(t-t_{\rm i})\rangle\langle V_0(t'-t_{\rm i})\rangle\right]. \tag{A77}$$

If we confine our attention to the simple case where the rate is constant, $r(t) = \bar{r}$, then the second term in brackets must be a constant, since $\langle V_0(t-t_{\rm i})\rangle$ involves averaging over all possible times $t_{\rm i}$, and with constant rate all these times are equally likely. So, if we don't worry about constants, we can write

$$\langle \delta V(t)\delta V(t')\rangle \sim \sum_{\rm i}\langle V_0(t-t_{\rm i})V_0(t'-t_{\rm i})\rangle \tag{A78}$$

$$= \left\langle \sum_{\rm i} V_0(t-t_{\rm i})V_0(t'-t_{\rm i})\right\rangle, \tag{A79}$$

and now we can use Eq (A66) to give

$$\langle \delta V(t)\delta V(t')\rangle = \bar{r}\int d\tau\, V_0(t-\tau)V_0(t'-\tau), \tag{A80}$$

where again we are neglecting a constant.

It is especially useful to convert the correlation function of voltage fluctuations into the corresponding power spectrum, since then any uncertainties about constants will go away.More precisely, if we had a constant term in the correlation function it would show up as a term $\sim \delta(\omega)$ in the power spectrum, and all we need to do is to be sure that we drop any such terms. In general, the power spectrum is

$$S_V(\omega) = \int d\tau\, e^{+i\omega\tau}\langle \delta V(t+\tau)\delta V(t)\rangle, \tag{A81}$$

and so in this case we have

$$
\begin{aligned}
S_V(\omega) &= \int d\tau\, e^{+i\omega\tau}\bar{r}\int d\tau'\, V_0(t+\tau-\tau')V_0(t-\tau') &\text{(A82)}\\
&= \bar{r}\int d\tau \int d\tau'\, e^{+i\omega\tau}e^{+i\omega(t-\tau')}V_0(t+\tau-\tau')e^{-i\omega(t-\tau')}V_0(t-\tau') &\text{(A83)}\\
&= \bar{r}\left[\int d\tau\, e^{+i\omega(\tau+t-\tau')}V_0(\tau+t-\tau')\right]\left[\int d\tau' e^{-i\omega(t-\tau')}V_0(t-\tau')\right] &\text{(A84)}\\
&= \bar{r}\left|\tilde{V}_0(\omega)\right|^2, &\text{(A85)}
\end{aligned}
$$

where in the last step we recognize the Fourier transform of the pulse shape $V_0(t)$. This is what we need for Eq (58) of the main text.

Problem 169: More carefully. Fill in the details of the calculation above, being sure to keep track of the floating constants. Verify that, when you are careful, there is no term $\sim \delta(\omega)$ in the power spectrum. Can you generalize this discussion to the case of time varying rates?

Portions of this section were adapted from Rieke et al (1997). The connection between power spectra and the shape of single photon (or more general Poisson) events is sometimes called Campell's theorem, and there is a classic discussion by Rice (1944–45), reprinted in the marvelous book edited by Wax (1954); the other articles in this book (by Chandrasekar and others) also are very much worth reading! Feynman & Hibbs (1965) give a beautiful discussion of how a Poisson stream of pulses comes to approximate continuous, Gaussian noise; of course there is much more in this book as well. For a more complete discussion of photon statistics, and the role of coherent states, one can look to yet another classic paper, Glauber (1963).

Feynman & Hibbs 1965: *Quantum Mechanics and Path Integrals.* RP Feynman & AR Hibbs (McGraw–Hill, New York, 1965).

Glauber 1963: Coherent and incoherent states of the radiation field. RJ Glauber, *Phys Rev* **131,** 2766–2788 (1963).

Rice 1944–45: Mathematical analysis of random noise. SO Rice, *Bell Sys Tech J* **23,** 282–332 (1944) & **24,** 46–156 (1945).

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

Wax 1954: *Selected Papers on Noise and Stochastic Processes.* N Wax, ed (Dover Publications, New York, 1954).

2. Correlations, power spectra and all that

Consider a function $x(t)$ that varies in time. We would like to describe a situation in which these variations are random, drawn out of some distribution. But now we need a distribution for a function, rather than for a finite set of variables. This shouldn't bother us, since such constructions are central to much of modern physics, for example in the path integral approach to quantum mechanics. We refer to distributions of functions as "distribution functionals" when we need to be precise.

One strategy for constructing distribution functionals is to start by discretizing time, so that we have at most a countable infinity of variables $x(t_1), x(t_2), x(t_3), \cdots$. Let's assume for simplicity that the mean value of x is zero. Then the first nontrivial characterization of the statistics of x is the covariance matrix,

$$C_{\rm ij} = \langle x(t_{\rm i})x(t_{\rm j})\rangle. \qquad \text{(A86)}$$

We recall that if a single variable y is drawn from a Gaussian distribution with zero mean, then we have

$$P(y) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{y^2}{2\sigma^2}\right]. \qquad \text{(A87)}$$

The generalization to multiple variables is

$$P(\{x_{\rm i}\}) = \frac{1}{\sqrt{(2\pi)^N \det C}}\exp\left[-\frac{1}{2}\sum_{\rm i,j=1}^{N} x_{\rm i}(C^{-1})_{\rm ij}x_{\rm j}\right], \qquad \text{(A88)}$$

where as usual det is the determinant and $(C^{-1})_{\rm ij}$ is the ij element of the matrix inverse to C; if we think of the $\{x_{\rm i}\}$ as a vector $\mathbf{x}$, then we can write, more compactly,

$$P(\{x_{\rm i}\}) = \frac{1}{\sqrt{(2\pi)^N \det C}}\exp\left[-\frac{1}{2}\mathbf{x}^T\cdot C^{-1}\cdot\mathbf{x}\right], \qquad \text{(A89)}$$

where $\mathbf{x}^T$ is the transpose of the vector $\mathbf{x}$. Just to be clear, this describes a Gaussian distribution, but we have no guarantee that $\mathbf{x}$ will be Gaussian.

Problem 170: Gaussian integrals. If you haven't done these before, now is a good time to check that the probability distribution

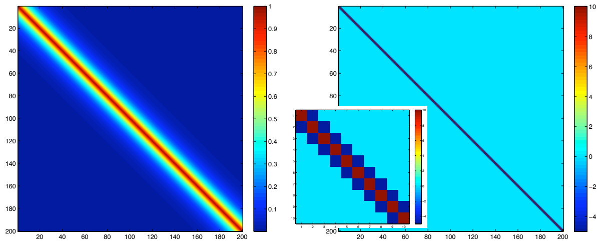

FIG. 167 Covariance matrix and its inverse. At left, the covariance matrix in Eq (A94), with $\Delta t/\tau_c = 0.1$. At right, the inverse matrix, with inset showing a 10×10 submatrix surrounding the diagonal.

in Eq (A89) is normalized. This requires you to show that

$$\int d^N x \, \exp\left[-\frac{1}{2}\mathbf{x}^T\cdot C^{-1}\cdot\mathbf{x}\right] = \sqrt{(2\pi)^N \det C}. \tag{A90}$$

While you're at it, you should also show that

$$\ln\det C = \mathrm{Tr}\ln C. \tag{A91}$$

This should be straightforward for the case which matters here, where C must have well defined, positive eigenvalues.

In general the covariance matrix C_{ij} can have an arbitrary structure, constrained only by symmetry and positivity of its eigenvalues. But when the index i refers to discrete time points, we have an extra constraint that comes from invariance under translations in time. Because there is no clock, we must have that

$$\langle x(t)x(t')\rangle = C_x(t-t'), \tag{A92}$$

with no dependence on the absolute time t or t'. As an example, if

$$C_x(t-t') = e^{-|t-t'|/\tau_c}, \tag{A93}$$

and $t_{\mathrm{n}} = n\Delta t$, then

$$C_{\mathrm{ij}} = \exp\left[-\left(\frac{\Delta t}{\tau_c}\right)|\mathrm{i}-\mathrm{j}|\right]. \tag{A94}$$

This is shown in Fig 167 for $\Delta t/\tau_c = 0.1$.

It is useful to look directly at the inverse matrix, also shown in Fig 167. We see that this inverse matrix consists almost entirely of zeros, except in the immediate neighborhood of the diagonal. This tell us that the inverse matrix actually is the discretization of a differential operator. Reflexively, seeing that we have to compute inverses and determinants of matrices, we should think about diagonalizing C. We recall from quantum mechanics that the eigenfunctions of an operator have to provide a representation of the underlying symmetries. In this case, the relevant symmetry is time translation, so we know to look at the Fourier functions, $e^{-i\omega t}$. In fact, once we have the hint that we should use a Fourier representation, we don't need the crutch of discrete time points any more. Let's see how this works.

We define the Fourier transform with the conventions

$$\tilde{x}(\omega) = \int_{-\infty}^{\infty} dt\, e^{+i\omega t}x(t), \tag{A95}$$

$$x(t) = \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} e^{-\omega t}\tilde{x}(\omega). \tag{A96}$$

Now if we compute the covariance of two frequency components, we have

$$\langle\tilde{x}(\omega)\tilde{x}(\omega')\rangle = \left\langle \int_{-\infty}^{\infty} dt\, e^{+i\omega t}x(t)\int_{-\infty}^{\infty} dt'\, e^{+i\omega' t'}x(t')\right\rangle \tag{A97}$$

$$= \int_{-\infty}^{\infty} dt\, e^{+i\omega t}\int_{-\infty}^{\infty} dt'\, e^{+i\omega' t'}\langle x(t)x(t')\rangle \tag{A98}$$

$$= \int_{-\infty}^{\infty} dt\, e^{+i\omega t}\int_{-\infty}^{\infty} dt'\, e^{+i\omega' t'}\int_{-\infty}^{\infty}\frac{d\Omega}{2\pi}e^{-i\Omega(t-t')}S_x(\Omega), \tag{A99}$$

where we introduce the Fourier transform of the correlation function,

$$S_x(\Omega) = \int_{-\infty}^{\infty} d\tau\, e^{+i\Omega\tau}C_x(\tau) \tag{A100}$$

$$C_x(t-t') = \int_{-\infty}^{\infty}\frac{d\Omega}{2\pi}e^{-i\Omega(t-t')}S_x(\Omega). \tag{A101}$$

Now we can rearrange the integrals in Eq (A99):

$$\begin{aligned}\langle\tilde{x}(\omega)\tilde{x}(\omega')\rangle &= \int_{-\infty}^{\infty} dt\, e^{+i\omega t}\int_{-\infty}^{\infty} dt'\, e^{+i\omega' t'}\int_{-\infty}^{\infty}\frac{d\Omega}{2\pi}e^{-i\Omega(t-t')}S_x(\Omega),\\ &= \int_{-\infty}^{\infty}\frac{d\Omega}{2\pi}S_x(\Omega)\left[\int_{-\infty}^{\infty} dt\, e^{i(\omega-\Omega)t}\right]\left[\int_{-\infty}^{\infty} dt'\, e^{i(\omega'+\Omega)t'}\right].\end{aligned} \tag{A102}$$

This is moment to recall the Fourier representation of the Dirac delta function. The delta function has the property that

$$\delta(z) = 0 \quad z \neq 0, \tag{A103}$$

$$\int dz\,\delta(z) = 1, \tag{A104}$$

if the domain of the integral includes $z = 0$. Then

$$\delta(z) = \int_{-\infty}^{\infty}\frac{dq}{2\pi}e^{-iqz}. \tag{A105}$$

Thus we recognize, in Eq (A102),

$$\int_{-\infty}^{\infty} dt\, e^{i(\omega-\Omega)t} = 2\pi\delta(\omega-\Omega), \tag{A106}$$

$$\int_{-\infty}^{\infty} dt'\, e^{i(\omega'+\Omega)t'} = 2\pi\delta(\omega'+\Omega). \tag{A107}$$

Substituting back into Eq (A102), we have

$$\begin{aligned}\langle\tilde{x}(\omega)\tilde{x}(\omega')\rangle &= \int_{-\infty}^{\infty}\frac{d\Omega}{2\pi}S_x(\Omega)\left[\int_{-\infty}^{\infty} dt\, e^{i(\omega-\Omega)t}\right]\left[\int_{-\infty}^{\infty} dt'\, e^{i(\omega'+\Omega)t'}\right].\\ &= \int_{-\infty}^{\infty}\frac{d\Omega}{2\pi}S_x(\Omega)2\pi\delta(\omega-\Omega)2\pi\delta(\omega'+\Omega) \qquad \text{(A108)}\\ &= S_x(\omega)2\pi\delta(\omega'+\omega). \qquad \text{(A109)}\end{aligned}$$

We see that, while different time points can be correlated with one another in complicated ways, the covariance of frequency components has a much simpler structure: $\tilde{x}(\omega)$ is correlated only with $\tilde{x}(-\omega)$.

This covariance structure, which couples positive and negative frequency components, makes sense when we realize that we are using a complex representation for real variables. To make a real variable $x(t)$, the Fourier transform must obey

$$\tilde{x}(-\omega) = \tilde{x}^*(\omega), \tag{A110}$$

so positive and negative frequency components are not independent—in fact they are redundant. It might be more natural to write Eq (A109) as

$$\langle\tilde{x}(\omega)\tilde{x}^*(\omega')\rangle = S_x(\omega)2\pi\delta(\omega'-\omega), \tag{A111}$$

making clear that frequency components are correlated with themselves, not with other frequencies.

We could instead think about the real and imaginary parts of the positive frequency components, which can be written as

$$\tilde{x}_{\mathrm{Re}}(\omega) = \frac{1}{2}\left[\tilde{x}(\omega)+\tilde{x}(-\omega)\right], \tag{A112}$$

and

$$\tilde{x}_{\mathrm{Im}}(\omega) = \frac{1}{2i}\left[\tilde{x}(\omega)-\tilde{x}(-\omega)\right]. \tag{A113}$$

With this representation, we can use the result in Eq (A109):

$$\begin{aligned}\langle\tilde{x}_{\mathrm{Re}}(\omega)\tilde{x}_{\mathrm{Re}}(\omega')\rangle &= \left\langle\frac{1}{2}\left[\tilde{x}(\omega)+\tilde{x}(-\omega)\right]\frac{1}{2}\left[\tilde{x}(\omega')+\tilde{x}(-\omega')\right]\right\rangle \qquad \text{(A114)}\\ &= \frac{1}{4}\left[\langle\tilde{x}(\omega)\tilde{x}(\omega')\rangle+\langle\tilde{x}(\omega)\tilde{x}(-\omega')\rangle+\langle\tilde{x}(-\omega)\tilde{x}(\omega')\rangle+\langle\tilde{x}(-\omega)\tilde{x}(-\omega')\rangle\right] \qquad \text{(A115)}\\ &= \frac{S_x(\omega)}{4}2\pi\left[\delta(\omega+\omega')+\delta(\omega-\omega')+\delta(-\omega+\omega')+\delta(-\omega-\omega')\right]. \qquad \text{(A116)}\end{aligned}$$

Because we are looking only at positive frequencies, $\omega+\omega'$ can never be zero, and hence the first and last delta functions can be dropped. The remaining two are actually the same, so we have

$$\langle \tilde{x}_{\rm Re}(\omega)\tilde{x}_{\rm Re}(\omega')\rangle = \frac{1}{2}S_x(\omega)2\pi\delta(\omega-\omega'). \tag{A117}$$

Similar calculations show that the imaginary parts of $\tilde{x}(\omega)$ have the same variance,

$$\begin{aligned}\langle \tilde{x}_{\rm Im}(\omega)\tilde{x}_{\rm Im}(\omega')\rangle &= \langle \tilde{x}_{\rm Re}(\omega)\tilde{x}_{\rm Re}(\omega')\rangle && \text{(A118)}\\ &= \frac{1}{2}S_x(\omega)2\pi\delta(\omega-\omega'), && \text{(A119)}\end{aligned}$$

while real and imaginary parts are uncorrelated,

$$\langle \tilde{x}_{\rm Re}(\omega)\tilde{x}_{\rm Im}(\omega')\rangle = 0. \tag{A120}$$

Problem 171: The other phase. Derive Eq's (A119) and (A120).

What does all this mean? We think of the random function of time $x(t)$ as being built out of frequency components, and each component has a real and imaginary part. The structure of the covariance matrix is such that different frequency components do not covary, and this makes sense—if we have covariation of different frequency components then we can beat them against each other to make a clock running at the difference frequency, and this would violate time translation invariance. Similarly, the fact that real and imaginary components do not covary means that there is no preferred phase, which again is consistent with (indeed, required by) time translation invariance.

We should be able to put these results on the covariance matrix together to describe the distribution functional for a Gaussian function of time. Since the real and imaginary parts are independent, let's start with just the real parts. We should have

$$P[\{\tilde{x}_{\rm Re}(\omega)\}] \propto \exp\left[-\frac{1}{2}\int_0^\infty\frac{d\omega}{2\pi}\int_0^\infty\frac{d\omega'}{2\pi}\tilde{x}_{\rm Re}(\omega)\mathcal{A}(\omega,\omega')\tilde{x}_{\rm Re}(\omega')\right], \tag{A121}$$

where $\mathcal{A}$ is the inverse of the covariance,

$$\int\frac{d\omega'}{2\pi}\mathcal{A}(\omega,\omega')\langle\tilde{x}_{\rm Re}(\omega')\tilde{x}_{\rm Re}(\omega'')\rangle = 2\pi\delta(\omega-\omega''). \tag{A122}$$

We can find $\mathcal{A}$ by substituting the explicit expression for the covariance and doing the integrals:

$$\begin{aligned}2\pi\delta(\omega-\omega'') &= \int\frac{d\omega'}{2\pi}\mathcal{A}(\omega,\omega')\langle\tilde{x}_{\rm Re}(\omega')\tilde{x}_{\rm Re}(\omega'')\rangle\\ &= \int\frac{d\omega'}{2\pi}\mathcal{A}(\omega,\omega')\frac{1}{2}S_x(\omega')2\pi\delta(\omega'-\omega'') && \text{(A123)}\\ &= \frac{1}{2}\mathcal{A}(\omega,\omega'')S_x(\omega''). && \text{(A124)}\end{aligned}$$

Thus, we have

$$\mathcal{A}(\omega,\omega'') = \frac{1}{S_x(\omega'')}4\pi\delta(\omega-\omega''). \tag{A125}$$

Substituting back into Eq (A121) for the probability distribution, we have

$$\begin{aligned}P[\{\tilde{x}_{\rm Re}(\omega)\}] &\propto \exp\left[-\frac{1}{2}\int_0^\infty\frac{d\omega}{2\pi}\int_0^\infty\frac{d\omega'}{2\pi}\tilde{x}_{\rm Re}(\omega)\mathcal{A}(\omega,\omega')\tilde{x}_{\rm Re}(\omega')\right], && \text{(A126)}\\ &= \exp\left[-\frac{1}{2}\int_0^\infty\frac{d\omega}{2\pi}\int_0^\infty\frac{d\omega'}{2\pi}\tilde{x}_{\rm Re}(\omega)\frac{4\pi\delta(\omega-\omega'')}{S_x(\omega'')}\tilde{x}_{\rm Re}(\omega')\right] && \text{(A127)}\\ &= \exp\left[-\int_0^\infty\frac{d\omega}{2\pi}\frac{\tilde{x}^2_{\rm Re}(\omega)}{S_x(\omega)}\right]. && \text{(A128)}\end{aligned}$$

Exactly the same argument applies to the imaginary parts of the Fourier components, and these are independent of the real parts, so we have

$$P[x(t)] = P[\{\tilde{x}_{\rm Re}(\omega), \tilde{x}_{\rm Im}(\omega)\}] \tag{A129}$$

$$\propto \exp\left[-\int_0^\infty \frac{d\omega}{2\pi}\frac{\tilde{x}^2_{\rm Re}(\omega)+\tilde{x}^2_{\rm Im}(\omega)}{S_x(\omega)}\right] \tag{A130}$$

$$= \frac{1}{Z}\exp\left[-\int_0^\infty \frac{d\omega}{2\pi}\frac{|\tilde{x}(\omega)|^2}{S_x(\omega)}\right] \tag{A131}$$

$$= \frac{1}{Z}\exp\left[-\frac{1}{2}\int_{-\infty}^\infty \frac{d\omega}{2\pi}\frac{|\tilde{x}(\omega)|^2}{S_x(\omega)}\right], \tag{A132}$$

where we have introduced the normalization constant Z.

It's useful to look at the example illustrated in Fig 167. Here we have $C_x(\tau) = \exp(-|\tau|/\tau_c)$, so the power spectrum is

$$S_x(\omega) = \int_{-\infty}^\infty d\tau\, e^{+i\omega\tau}e^{-|\tau|/\tau_c} \tag{A133}$$

$$= \int_{-\infty}^0 d\tau\, e^{(+i\omega+1/\tau_c)\tau} + \int_0^\infty d\tau\, e^{(+i\omega-1/\tau_c)\tau} \tag{A134}$$

$$= \frac{1}{(+i\omega+1/\tau_c)} + \frac{1}{-(+i\omega-1/\tau_c)} \tag{A135}$$

$$= \frac{\tau_c}{1+i\omega\tau_c} + \frac{\tau_c}{1-i\omega\tau_c} \tag{A136}$$

$$= \frac{2\tau_c}{1+(\omega\tau_c)^2}. \tag{A137}$$

This means that the probability distribution functional has the form

$$P[x(t)] = \frac{1}{Z}\exp\left[-\frac{1}{2}\int_{-\infty}^\infty \frac{d\omega}{2\pi}\frac{|\tilde{x}(\omega)|^2}{S_x(\omega)}\right]$$
$$= \frac{1}{Z}\exp\left[-\frac{1}{4\tau_c}\int_{-\infty}^\infty \frac{d\omega}{2\pi}[1+(\omega\tau_c)^2]|\tilde{x}(\omega)|^2\right]. \tag{A138}$$

We recall that

$$\int_{-\infty}^\infty \frac{d\omega}{2\pi}|\tilde{x}(\omega)|^2 = \int dt\, x^2(t). \tag{A139}$$

More subtly,

$$\int_{-\infty}^\infty \frac{d\omega}{2\pi}(\omega\tau_c)^2|\tilde{x}(\omega)|^2 = \tau_c^2\int_{-\infty}^\infty \frac{d\omega}{2\pi}|-i\omega\tilde{x}(\omega)|^2 \tag{A140}$$

$$= \tau_c^2\int dt\left[\frac{dx(t)}{dt}\right]^2, \tag{A141}$$

where we recognize $-i\omega\tilde{x}(\omega)$ as the Fourier transform of $dx(t)/dt$. Thus we can write

$$P[x(t)] = \frac{1}{Z}\exp\left[-\frac{1}{4\tau_c}\int_{-\infty}^\infty \frac{d\omega}{2\pi}[1+(\omega\tau_c)^2]|\tilde{x}(\omega)|^2\right]$$
$$= \frac{1}{Z}\exp\left[-\frac{1}{4\tau_c}\int dt\left(\tau_c^2\dot{x}^2(t)+x^2(t)\right)\right] \tag{A142}$$

This shows explicitly, as promised above, that inverting the covariance matrix gives rise to differential operators. This example also is nice because it produces a probability distribution functional for trajectories $x(t)$ that reminds us of a (Euclidean) path integral in quantum mechanics, in this case for the harmonic oscillator.

Let's push a little further and see if we can evaluate the normalization constant Z. By definition, we have

$$Z = \int \mathcal{D}x\,\exp\left[-\frac{1}{4\tau_c}\int dt\left(\tau_c^2\dot{x}^2(t)+x^2(t)\right)\right], \tag{A143}$$

where $\int\mathcal{D}x$ denotes an integral over all the functions $x(t)$. We have the general result for an N dimensional Gaussian integral,

$$\int d^Nx\,\exp\left[-\frac{1}{2}\mathbf{x}^T\cdot\hat{A}\cdot\mathbf{x}\right] = \sqrt{\frac{(2\pi)^N}{\det\hat{A}}} \tag{A144}$$

$$= \sqrt{(2\pi)^N}\exp\left[-\frac{1}{2}{\rm Tr}\ln\hat{A}\right], \tag{A145}$$

where $\hat{A}$ is a matrix. Here we need to let the number of dimensions become infinite, since we are integrating over functions. As you may recall from discussions of the path integral in quantum mechanics, there is some arbitrariness about how we do this, or, more formally, in how we define the measure $\mathcal{D}x$. A fairly standard choice is to absorb the $\sqrt{2\pi}$, so that, in the time window $0 < t < T$,

$$\mathcal{D}x = \lim_{dt\to 0}\prod_{\rm n=0}^{T/dt}\frac{dx(t_{\rm n})}{\sqrt{2\pi}}, \qquad t_{\rm n} = {\rm n}\cdot dt. \tag{A146}$$

Notice that before we send $dt \to 0$, we have an integral over a finite number of points, so we should be able to carry over the results we know, and just interpret the limits correctly.

The Gaussian functional integrals that we want to do have the general form

$$\int\mathcal{D}x\,\exp\left[-\frac{1}{2}\int dt\int dt' x(t)\hat{K}(t,t')x(t')\right],$$

where $\hat{K}$ is an operator. Carrying over what we know from the case of finite matrices [Eq (A145)], we have

$$\int \mathcal{D}x \, \exp\left[-\frac{1}{2}\int dt \int dt' x(t)\hat{K}(t,t')x(t')\right] = \exp\left[-\frac{1}{2}\mathrm{Tr}\ln\hat{K}\right]. \tag{A147}$$

Our only problem is to say what we mean by $\mathrm{Tr}\ln\hat{K}$. Since $\hat{K}$ is an operator, we can ask for its spectrum, that is the eigenvalues and eigenfunctions. This means that we need to solve the equations

$$\int_0^T dt'\hat{K}(t,t')u_\mu(t') = \Lambda_\mu u_\mu(t), \tag{A148}$$

where we are careful here to note that we are working in window $0 < t < T$. In the basis formed by the eigenfunctions, of course $\hat{K}$ is diagonal. As with matrices, when an operator is diagonal we can take the log element by element, and then computing the trace requires us to sum over these diagonal elements; recall that traces and determinants are invariant, se we can use this convenient basis and not worry about generality. Thus,

$$\mathrm{Tr}\ln\hat{K} = \sum_\mu \ln\Lambda_\mu. \tag{A149}$$

How does this work for our case? First, we need to identify the operator $\hat{K}$. In the exponential of $P[x(t)]$ we have

$$\int dt\left[\frac{\tau_c}{2}\left(\frac{dx(t)}{dt}\right)^2 + \frac{1}{2\tau_c}x^2(t)\right].$$

To get this into a more standard form we need to integrate by parts,

$$\int dt\left[\frac{\tau_c}{2}\left(\frac{dx(t)}{dt}\right)^2 + \frac{1}{2\tau_c}x^2(t)\right] = \int dt\, x(t)\left[-\frac{\tau_c}{2}\frac{d^2}{dt^2} + \frac{1}{2\tau_c}\right]x(t). \tag{A150}$$

This allows us to identify We now see that our integral for Z in Eq (A143) can we written

$$\hat{K}(t',t) = \delta(t'-t)\left[-\frac{\tau_c}{2}\frac{d^2}{dt^2} + \frac{1}{2\tau_c}\right]. \tag{A151}$$

This is a linear operator, and also time translation invariant (again). So we know that the eigenfunctions are $e^{-i\omega t}$, and since we are in a finite window of duration T we should use only those frequency components that 'fit' into the window, $\omega_\mathrm{n} = 2\pi\mathrm{n}/T$ for integer n. We have

$$\int_0^T dt\,\delta(t'-t)\left[-\frac{\tau_c}{2}\frac{d^2}{dt^2} + \frac{1}{2\tau_c}\right]e^{-i\omega_\mathrm{n}t} = \left(\frac{\tau_c\omega_\mathrm{n}^2}{2} + \frac{1}{2\tau_c}\right)e^{-i\omega_\mathrm{n}t'}, \tag{A152}$$

so that the eigenvalues are

$$\Lambda(\omega_\mathrm{n}) = \left(\frac{\tau_c\omega_\mathrm{n}^2}{2} + \frac{1}{2\tau_c}\right) = \frac{1+(\omega_\mathrm{n}\tau_c)^2}{2\tau_c}. \tag{A153}$$

Notice that these are just the inverses of the power spectrum,

$$\Lambda(\omega_\mathrm{n}) = \frac{1}{S_x(\omega_\mathrm{n})}. \tag{A154}$$

This makes sense, of course, when we look back at Eq (A132).

To finish the calculation, we have

$$Z = \exp\left[-\frac{1}{2}\sum_\mu \Lambda_\mu\right] \tag{A155}$$

$$= \exp\left[-\frac{1}{2}\sum_\mathrm{n}\ln\left(\frac{1}{S_x(\omega_\mathrm{n})}\right)\right] \tag{A156}$$

$$= \exp\left[\frac{1}{2}\sum_\mathrm{n}\ln S_x(\omega_\mathrm{n})\right]. \tag{A157}$$

Finally, we need to do the sum. As the time window T becomes large, the spacing between frequency components, $\Delta\omega = 2\pi/T$, become small, and we expect that the sum approaches an integral.[91] Thus, for any function of $\omega_{\rm n}$,

$$\sum_{\rm n} f(\omega_{\rm n}) = \frac{1}{\Delta\omega}\sum_{\rm n}\Delta\omega f(\omega_{\rm n}) \qquad (A159)$$

$$\to \frac{1}{\Delta\omega}\int d\omega\, f(\omega) \qquad (A160)$$

$$= T\int\frac{d\omega}{2\pi}f(\omega). \qquad (A161)$$

At last, this gives us

$$Z = \exp\left[\frac{T}{2}\int\frac{d\omega}{2\pi}\ln S_x(\omega)\right]. \qquad (A162)$$

Putting the pieces together, we have the probability distribution functional for a Gaussian $x(t)$,

$$P[x(t)] = \exp\left[+\frac{T}{2}\int_{-\infty}^{\infty}\ln S_x(\omega) - \frac{1}{2}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\frac{|\tilde{x}(\omega)|^2}{S_x(\omega)}\right]. \qquad (A163)$$

Not every case we look at will be Gaussian, but this helps to get us started.

Problem 172: Generality. We made an effort to evaluate Z in the specific case where $C_x(\tau) = e^{-|\tau|/\tau_c}$, but we wrote the final result in a very general form, Eq (A163). Show that this slide into generality was justified.

Problem 173: Nonzero means and signal to noise ratios. We should be able to carry everything through in the case where the mean $x(t)$ is not zero. For example, if we just have background noise described by some spectrum $\mathcal{N}(\omega)$, then

$$P_{\rm noise}[x(t)] = \exp\left[+\frac{T}{2}\int_{-\infty}^{\infty}\ln\mathcal{N}(\omega) - \frac{1}{2}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\frac{|\tilde{x}(\omega)|^2}{\mathcal{N}(\omega)}\right]. \qquad (A164)$$

If there is an added signal $x_0(t)$, the distribution functional becomes

$$P_{\rm signal}[x(t)] = \exp\left[+\frac{T}{2}\int_{-\infty}^{\infty}\ln\mathcal{N}(\omega) - \frac{1}{2}\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\frac{|\tilde{x}(\omega) - \tilde{x}_0(\omega)|^2}{\mathcal{N}(\omega)}\right]. \qquad (A165)$$

Suppose that you observe some particular $x(t)$, and you have to decide whether this came from the signal or noise distribution, that is, you have to decide whether the signal was present; for simplicity assume that the two possibilities are equally likely a priori. As discussed in Chapter 1, to make such decisions optimally you should use the relative probabilities that the signal or noise could give rise to your data. In particular, consider computing the "log likelihood ratio,"

$$\lambda[x(t)] \equiv \ln\left(\frac{P_{\rm signal}[x(t)]}{P_{\rm noise}[x(t)]}\right) \qquad (A166)$$

(a.) Give a simple expression for $\lambda[x(t)]$. Show that it is a linear functional of $x(t)$.

(b.) Show that, when the $x(t)$ are drawn at random out of either $P_{\rm signal}$ or $P_{\rm noise}$, $\lambda[x(t)]$ is a Gaussian random variable. Find the means, $\langle\lambda\rangle_{\rm noise}$ and $\langle\lambda\rangle_{\rm signal}$, and the variances $\langle(\delta\lambda)^2\rangle_{\rm noise}$ and $\langle(\delta\lambda)^2\rangle_{\rm signal}$, in the two distributions. Hint: you should see that $\langle(\delta\lambda)^2\rangle_{\rm noise} = \langle(\delta\lambda)^2\rangle_{\rm signal}$.

(c.) Sketch the distributions $P_{\rm noise}(\lambda)$ and $P_{\rm signal}(\lambda)$. Show that your ability to make reliable discriminations is determined only by the signal to noise ratio,

$$SNR = \frac{\left(\langle\lambda\rangle_{\rm signal} - \langle\lambda\rangle_{\rm noise}\right)^2}{\langle(\delta\lambda)^2\rangle}, \qquad (A167)$$

and that we can write

$$SNR = \int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\frac{|\tilde{x}_0(\omega)|^2}{\mathcal{N}(\omega)}. \qquad (A168)$$

(d.) In rod cells, a single photon produces a current pulse with the approximate form $x_0(t) = I_1(t/\tau)^3e^{-t/\tau}$. The power spectrum of continuous background noise is approximately $\mathcal{N}(\omega) = A/[1+(\omega\tau)^2]^2$, with the same value of τ. Evaluate the peak current, $I_{\rm peak}$, and total variance of the background noise, σ_I^2. A naive estimate of the signal to noise ratio is just $SNR_{\rm naive} = (I_{\rm peak}/\sigma_I)^2$. Show that the optimal signal to noise ratio, computed from Eq (A168), is larger. Why?

Is there anything more to say here? Maybe some discussion of "states" of molecules and correlation functions? Perhaps some references.

3. Electronic transition in large molecules

In this section we'll outline an honest calculation that reproduces the intuition of Figs 19 and 20. We have a system with two electronic states, which we can represent as a spin one–half; let spin down be the ground state and spin up be the excited state. The Born–Oppenheimer approximation tells us that we can think of the atoms in a molecule as moving in a potential determined by the electronic state,[92] which we denote by $V_\uparrow(\mathbf{q})$ and $V_\downarrow(\mathbf{q})$ in the excited and ground states, respectively; $\mathbf{q}$ stands for all the atomic coordinates (not just the one in the sketches above). Since we are observing photon absorption, there must be a matrix element that connects the two electronic states and couples to the electromagnetic field; we'll assume that, absent symmetries, this coupling is dominated by an electric dipole term. In principle the dipole matrix element $\vec{d}$ could depend upon the atomic coordinates, but we'll neglect this effect.[93] Putting the pieces together, we have the Hamiltonian for the molecule

$$\mathbf{H} = \mathbf{K} + \frac{1}{2}(1+\sigma_z)V_\uparrow(\mathbf{q}) + \frac{1}{2}(1-\sigma_z)V_\downarrow(\mathbf{q}) + \vec{d}\cdot\vec{E}(\sigma_+ + \sigma_-), \qquad (A169)$$

[91] There is an analogous result for summing over the states of particles in a box in quantum systems; recall that the states are labelled by their wavevector $\mathbf{k}$, and in three dimensions we have

$$\sum_{\mathbf{k}} \to V\int\frac{d^3k}{(2\pi)^3}, \qquad (A158)$$

where V is the volume of the box.

[92] As in the main text, I'll use "atoms" and "nuclei" interchangeably.

[93] In practice, this is a small effect. You should think about why this is true.

where $\mathbf{K}$ is the kinetic energy of the atoms. To this we should of course add the usual Hamiltonian for the electromagnetic field.

We are interested in computing the rate at which photons of energy $\hbar\Omega$ are absorbed, and of course we will do this as a perturbation expansion in the term $\sim \vec{d}$. The result of such a calculation can be presented as the 'Golden rule' for transition rates, but this formulation hides the underlying dynamics. So, at the risk of being pedantic, I'll go through the steps that usually lead to the Golden rule and take a detour that leads us to a formula in which the dynamics of atomic motions are more explicit.[94]

We start our system in the ground state of the electrons ($|\downarrow\rangle$), in some initial state ($|i\rangle$) of the atomic coordinates, and in the presence of one photon of wavevector $\vec{k}$ and frequency $\Omega = c|\vec{k}|$ (polarization is an unnecessary complication here). As the system evolves under the Hamiltonian $\mathbf{H}$, at some time t we want to measure the probability of finding the system in the excited state $|\uparrow\rangle$, in some other state of the atoms $|f\rangle$, and absent the photon. The general statement is that quantum states evolve as

$$|\psi(0)\rangle \rightarrow |\psi(t)\rangle = \mathbf{T}\exp\left[-\frac{i}{\hbar}\int_0^t d\tau \mathbf{H}(\tau)\right]|\psi(0)\rangle, \tag{A170}$$

where $\mathbf{T}$ is the time ordering operator. Thus, for our particular problem, the probability of starting in state $|\downarrow, i, \vec{k}\rangle$ and ending in state $|\uparrow, f, \emptyset\rangle$ is given by

$$p_{i\rightarrow f}(t) = \left|\langle\emptyset, f, \uparrow|\mathbf{T}\exp\left[-\frac{i}{\hbar}\int_0^t d\tau \mathbf{H}(\tau)\right]|\downarrow, i, \vec{k}\rangle\right|^2. \tag{A171}$$

In fact, we don't care about the final state of the atoms, and we can't select their initial state—this comes out of the Boltzmann distribution. So we really should compute

$$P(t) = \sum_{i,f}\left|\langle\emptyset, f, \uparrow|\mathbf{T}\exp\left[-\frac{i}{\hbar}\int_0^t d\tau \mathbf{H}(\tau)\right]|\downarrow, i, \vec{k}\rangle\right|^2 p_i, \tag{A172}$$

where p_i is the probability of being in the initial atomic state i.

As usual, we will break the Hamiltonian into two pieces, $\mathbf{H} = \mathbf{H}_0 + \mathbf{H}_1$, and do perturbation theory in $\mathbf{H}_1$. We choose $\mathbf{H}_1 = \vec{d}\cdot\vec{E}(\sigma_+ + \sigma_-)$, which is the only term that connects the states $|\downarrow\rangle$ and $|\uparrow\rangle$. The leading term in the perturbation series thus becomes

$$P(t) \approx \frac{1}{\hbar^2}\sum_{i,f}\left|\langle\emptyset, f, \uparrow|\mathbf{T}e^{-\frac{i}{\hbar}\int_0^t d\tau \mathbf{H}_0(\tau)}\int_0^t d\tau' \mathbf{H}_1(\tau')|\downarrow, i, \vec{k}\rangle\right|^2 p_i. \tag{A173}$$

If we look more carefully at the amplitude, we have

$$\langle\emptyset, f, \uparrow|\mathbf{T}e^{-\frac{i}{\hbar}\int_0^t d\tau \mathbf{H}_0(\tau)}\int_0^t d\tau' \mathbf{H}_1(\tau')|\downarrow, i, \vec{k}\rangle = \int_0^t d\tau' \langle f|\mathbf{T}\left(e^{-\frac{i}{\hbar}\int_{\tau'}^t d\tau \mathbf{H}_\uparrow(\tau)}\right)\vec{d}\cdot\langle\emptyset|\vec{E}|\vec{k}\rangle\mathbf{T}\left(e^{-\frac{i}{\hbar}\int_0^{\tau'} d\tau \mathbf{H}_\downarrow(\tau)}\right)|i\rangle e^{-i\Omega\tau'}, \tag{A174}$$

where τ' is the moment at which the term $\mathbf{H}_1 \sim \sigma_+$ acts to flip the state from $|\downarrow\rangle$ to $|\uparrow\rangle$; the terms $\mathbf{H}_{\downarrow,\uparrow}$ are defined by

$$\mathbf{H}_\downarrow = \mathbf{K} + V_\downarrow(\mathbf{q}), \tag{A175}$$

and similarly for $\mathbf{H}_\uparrow$. The key point is that when we square this amplitude and sum over final states, we can identify this as a sum over a complete set of states, and we recall that

$$\sum_f |f\rangle\langle f| = \mathbf{1}, \tag{A176}$$

the unit operator. Further, to keep things simple, let's assume that motion of the atoms is approximately classical. Because the terms $\mathbf{H}_{\uparrow,\downarrow}$ depend only on the atomic coordinates and momenta, the classical approximation means that we don't have to worry about the non–commutativity of these operators at different times, and we can drop the formalities of time ordering. Putting all of the terms together, we can rewrite $P(t)$ from Eq (A173):

$$P(t) \approx \frac{(\vec{d}\cdot\langle\emptyset|\vec{E}|\vec{k}\rangle)^2}{\hbar^2}\int_0^t d\tau_1 \int_0^t d\tau_2\, e^{+i\Omega(\tau_1-\tau_2)} \sum_i p_i \langle i| e^{+\frac{i}{\hbar}\int_0^{\tau_1} d\tau \mathbf{H}_\downarrow(\tau)} e^{+\frac{i}{\hbar}\int_{\tau_1}^{t} d\tau \mathbf{H}_\uparrow(\tau)} e^{-\frac{i}{\hbar}\int_{\tau_2}^{t} d\tau \mathbf{H}_\uparrow(\tau)} e^{-\frac{i}{\hbar}\int_0^{\tau_2} d\tau \mathbf{H}_\downarrow(\tau)} |i\rangle \tag{A177}$$

$$= \frac{(\vec{d}\cdot\langle\emptyset|\vec{E}|\vec{k}\rangle)^2}{\hbar^2}\int_0^t d\tau_1 \int_0^t d\tau_2\, e^{+i\Omega(\tau_1-\tau_2)} \sum_i p_i \langle i| \exp\left(+\frac{i}{\hbar}\int_{\tau_1}^{\tau_2} d\tau [\mathbf{H}_\uparrow(\tau) - \mathbf{H}_\downarrow(\tau)]\right)|i\rangle \tag{A178}$$

$$\propto \int_0^t d\tau_1 \int_0^t d\tau_2 e^{+i\Omega(\tau_1-\tau_2)} \left\langle \exp\left[+\frac{i}{\hbar}\int_{\tau_1}^{\tau_2} d\tau\, \epsilon[\mathbf{q}(\tau)]\right]\right\rangle, \tag{A179}$$

where $\epsilon = \mathbf{H}_\uparrow - \mathbf{H}_\downarrow = V_\uparrow - V_\downarrow$ is the instantaneous energy difference between the ground and excited states, which fluctuates as the atomic coordinates fluctuate, and $\langle\cdots\rangle$ denotes and average over these fluctuations.

Problem 174: Missing steps. Fill in the steps leading to Eq (A179). If you are more ambitious, try the case where the atomic motions are fully quantum mechanical.

Notice that the integrand in Eq (A179) depends only on the time difference $\tau_2 - \tau_1$. Thus, we are doing an integral of the form

$$\int_0^t d\tau_1 \int_0^t d\tau_2\, F(\tau_2 - \tau_1). \tag{A180}$$

It seems natural to rewrite this integral over the (τ_1, τ_2) plane in terms of an integral over the time difference and the mean. In the limit that t is large, this yields

$$\int_0^t d\tau_1 \int_0^t d\tau_2\, F(\tau_2 - \tau_1) \rightarrow t\int_{-\infty}^{\infty} d\tau\, F(\tau). \tag{A181}$$

Thus, we have

$$P(t) \propto t\int_{-\infty}^{\infty} d\tau\, e^{+i\Omega\tau}\left\langle \exp\left[-\frac{i}{\hbar}\int_0^{\tau} d\tau'\, \epsilon[\mathbf{q}(\tau')]\right]\right\rangle, \tag{A182}$$

so that the transition rate or absorption cross–section for photons of frequency Ω becomes

$$\sigma(\Omega) \propto \left\langle \int_{-\infty}^{\infty} d\tau\, \exp\left[+i\Omega\tau - \frac{i}{\hbar}\int_0^{\tau} d\tau'\, \epsilon[\mathbf{q}(\tau')]\right]\right\rangle. \tag{A183}$$

Now we can recover the intuition of Fig 19 as a saddle point approximation to the integral in Eq (A183). We recall that the saddle point approximation is

$$\int dt\, \exp\left[+i\phi(t)\right] \approx \sqrt{\frac{2\pi}{|\phi''(t_*)|}}\exp\left[+i\phi(t_*)\right], \tag{A184}$$

where the time t_* is defined by

$$\left.\frac{d\phi(t)}{dt}\right|_{t=t_*} = 0. \tag{A185}$$

The condition for validity of the approximation is that the time scale

$$\delta t \sim 1/\sqrt{|\phi''(t_*)|} \tag{A186}$$

be small compared with the intrinsic time scales for variation of $\phi(t)$. As applied to Eq (A183), the saddle point condition is

$$0 = \left.\frac{d}{d\tau}\left[+i\Omega\tau - \frac{i}{\hbar}\int_0^{\tau} d\tau'\, \epsilon[\mathbf{q}(\tau')]\right]\right|_{\tau=\tau_*} \tag{A187}$$

$$= i\Omega - \frac{i}{\hbar}\epsilon[\mathbf{q}(\tau_*)] \tag{A188}$$

$$\hbar\Omega = \epsilon[\mathbf{q}(\tau_*)]. \tag{A189}$$

Thus, the saddle point condition states that the integral defining the cross–section is dominated by moments when the instantaneous difference between the ground and excited state energies matches the photon energy. But this instantaneous difference $\epsilon[\mathbf{q}]$ is exactly the ‘vertical’ energy difference in Fig 19. Since this integral is inside an expectation value over the fluctuations in atomic coordinates, the cross–section will be proportional to the probability that this matching condition is obeyed.

If the sketch in Fig 19 is equivalent to a saddle point approximation, we have to consider conditions for validity of this approximation. The time scale defined by Eq (A186) becomes

$$\delta t \sim \left|\frac{1}{\hbar}\frac{d\epsilon[\mathbf{q}(\tau)]}{d\tau}\right|^{-1/2} \sim \sqrt{\frac{\hbar}{\epsilon' v}}, \tag{A190}$$

where ϵ' is the slope of the energy difference as a function of atomic coordinates, and v is a typical velocity for motion along these coordinates. Thus large slopes result in smaller values of δt, and of course this time scales as $\sqrt{\hbar}$. The natural time scale of motion along the atomic coordinates is given by vibrational periods, or $\omega_{\rm vib}^{-1} = \tau_{\rm vib} \sim \Delta/v$, where Q is a typical displacement from equilibrium. This lets us write

$$\delta t \sim \sqrt{\frac{\hbar}{\epsilon' v}} \sim \sqrt{\frac{\hbar\omega_{\rm vib}}{\epsilon' Q}\cdot\frac{Q}{v\omega_{\rm vib}}} \sim \tau_{\rm vib}\sqrt{\frac{\hbar\omega_{\rm vib}}{\epsilon' Q}}. \quad \text{(A191)}$$

We see that $\delta t \ll \tau_{\rm vib}$ if the energy $\epsilon' Q$ is much larger than the energy of vibrational quanta $\hbar\omega_{\rm vib}$. But $\epsilon' Q$ is the range of energy differences between the ground and excited states that the molecule can access as it fluctuates—and this is the width of the absorption spectrum. Thus, self–consistently, if we find that the width of the spectrum is large compared to the vibrational quanta, then our saddle point approximation is accurate.

We can go a bit further if we specialize to the case where, as in Fig 20, the different potential surfaces are exactly Hookean springs, that is when the dynamics of atomic motions are harmonic oscillators. In the general case there are many normal modes, so we would write

$$V_\uparrow(\mathbf{q}) = \frac{1}{2}\sum_{\rm i}\omega_{\rm i}^2 q_{\rm i}^2 \quad \text{(A192)}$$

$$V_\uparrow(\mathbf{q}) = \epsilon_0 + \frac{1}{2}\sum_{\rm i}\omega_{\rm i}^2 (q_{\rm i} - \Delta_{\rm i})^2. \quad \text{(A193)}$$

In this case,

$$\begin{aligned}\epsilon[\mathbf{q}(t)] &\equiv V_\uparrow[\mathbf{q}(t)] - V_\uparrow[\mathbf{q}(t)] && \text{(A194)}\\ &= \epsilon_0 + \frac{1}{2}\sum_{\rm i}\omega_{\rm i}^2\Delta_{\rm i}^2 - \sum_{\rm i}\omega_{\rm i}^2\Delta_{\rm i} q_{\rm i}(t) && \text{(A195)}\\ &= \hbar\Omega_{\rm peak} - X(t), && \text{(A196)}\end{aligned}$$

where the generalized coordinate $X(t)$ is given by a weighted combination of all the modes,

$$X(t) = \sum_{\rm i}\omega_{\rm i}^2\Delta_{\rm i} q_{\rm i}(t). \quad \text{(A197)}$$

Equation (A183) for the absorption cross–section thus becomes

$$\begin{aligned}\sigma(\Omega) &\propto \left\langle \int_{-\infty}^{\infty} d\tau\, \exp\left[+i\Omega\tau - \frac{i}{\hbar}\int_0^\tau d\tau'\, \epsilon[\mathbf{q}(\tau')]\right]\right\rangle\\ &= \left\langle \int_{-\infty}^{\infty} d\tau\, \exp\left[+i\Omega\tau - \frac{i}{\hbar}\int_0^\tau d\tau'\, (\hbar\Omega_{\rm peak} - X(\tau'))\right]\right\rangle\\ &= \int_{-\infty}^{\infty} d\tau\, e^{+i(\Omega-\Omega_{\rm peak})\tau}\left\langle \exp\left[+\frac{i}{\hbar}\int_0^\tau d\tau' X(\tau')\right]\right\rangle. \quad \text{(A198)}\end{aligned}$$

The key point is that, because $X(t)$ is a sum of harmonic oscillator coordinates, its fluctuations are drawn from a Gaussian distribution when we compute the average $\langle\cdots\rangle$ over the equilibrium ensemble.

Problem 175: Gaussian averages. Derive Eq (A199).

We recall that, if y is a Gaussian random variable, then

$$\langle e^y\rangle = \exp\left[\langle y\rangle + \frac{1}{2}\langle(\delta y)^2\rangle\right]. \quad \text{(A199)}$$

In the present case, the role of y is played by an integral over the trajectory of $X(t)$, but this shouldn't bother us:

$$\begin{aligned}&\left\langle \exp\left[+\frac{i}{\hbar}\int_0^\tau d\tau' X(\tau')\right]\right\rangle\\ &\quad = \exp\left[\frac{1}{2}\left\langle\left(\frac{i}{\hbar}\int_0^\tau d\tau' X(\tau')\right)^2\right\rangle\right] && \text{(A200)}\\ &\quad = \exp\left[-\frac{1}{2\hbar^2}\int_0^\tau d\tau_1\int_0^\tau d\tau_2 \langle X(\tau_1)X(\tau_2)\rangle\right], && \text{(A201)}\end{aligned}$$

where we start by making use of the fact that $\langle X\rangle = 0$.

We see from Eq (A201) that the shape of the absorption spectrum is determined by the correlation function of the modes to which the electronic transition are coupled, that is $C_X(\tau_1 - \tau_2) = \langle X(\tau_1)X(\tau_2)\rangle$ If these modes have relatively slow dynamics, then the time scales τ that enter the integral we need to do will be much shorter than the time scales over which this correlation function varies. In this limit we can approximate

$$\int_0^\tau d\tau_1 \int_0^\tau d\tau_2 \langle X(\tau_1)X(\tau_2)\rangle \approx \int_0^\tau d\tau_1 \int_0^\tau d\tau_2 \langle X(0)X(0)\rangle \\ = \langle X^2\rangle \tau^2. \quad \text{(A202)}$$

Notice also that

$$\langle X^2\rangle = \left\langle \left(\sum_{\rm i} \omega_{\rm i}^2 \Delta_{\rm i} q_{\rm i}\right)^2 \right\rangle = \sum_{\rm i} \omega_{\rm i}^4 \Delta_{\rm i}^2 \langle q_{\rm i}^2\rangle; \quad \text{(A203)}$$

in the classical limit we have $\langle q_{\rm i}^2\rangle = k_BT/\omega_{\rm i}^2$, and hence

$$\langle X^2\rangle = k_BT \sum_{\rm i} \omega_{\rm i}^2 \Delta_i^2 = 2k_BT\lambda, \quad \text{(A204)}$$

where λ generalizes the reorganization energy or Stokes' shift to the case of many modes. Finally, putting these pieces together, we have

$$\sigma(\Omega) \propto \int_{-\infty}^{\infty} d\tau \exp\left[+i(\Omega - \Omega_{\rm peak})\tau\right] \\ \times \exp\left[-\frac{1}{2\hbar^2}\int_0^\tau d\tau_1 \int_0^\tau d\tau_2 \langle X(\tau_1)X(\tau_2)\rangle\right] \\ \text{(A205)}$$

$$\approx \int_{-\infty}^{\infty} d\tau \,\exp\left[+i(\Omega - \Omega_{\rm peak})\tau - \frac{\tau^2\lambda k_BT}{\hbar^2}\right] \text{(A206)}$$

$$= \sqrt{\frac{\pi\hbar^2}{\lambda k_BT}} \exp\left[-\frac{(\hbar\Omega - \hbar\Omega_{\rm peak})^2}{4\lambda k_BT}\right]. \quad \text{(A207)}$$

This result should look familiar from Eq (66).

The calculation we have done here also allows us to look more precisely at the limits to our approximation. The integral in Eq (A206) is a Gaussian integral over τ, which means that it is done exactly by the saddle point method. The characteristic time which emerges from this is

$$\delta t \sim \frac{\hbar}{\sqrt{\lambda k_BT}}. \quad \text{(A208)}$$

If the typical vibrational time scales that enter into $C_X(\tau)$ are $\tau_{\rm vib} \sim \hbar/k_BT$, then the condition for validity of our approximation becomes $\lambda \gg k_BT$. Tracing the factors through, our approximate result should be valid if the predicted width of the absorption spectrum is (in energy units) larger than k_BT, or roughly one percent of $\hbar\Omega_{\rm peak}$. This is a rather gentle condition, suggesting that whenever the model of harmonic normal modes is correct, something like the saddle point approximation ought to work.

In fact, this calculation also gives us insight into another way that our semi–classical intuition from Fig 19 can fail. If, for example, there was just a single normal mode, we would have $X = gq(t)$, where $g = \omega^2\Delta$. But if there is just this one mode, and no other degrees of freedom to suck energy out of this mode, we must have

$$\langle q(t)q(t')\rangle = \frac{k_BT}{\omega^2}\cos[\omega(t-t')], \quad \text{(A209)}$$

so the integral [Eq (A205)] which defines the cross–section becomes

$$\sigma(\Omega) \propto \int_{-\infty}^{\infty} d\tau \,\exp\left[+i(\Omega - \Omega_{\rm peak})\tau - \frac{\Delta^2\omega^2 k_BT}{2\hbar^2}\int_0^\tau d\tau_1 \int_0^\tau d\tau_2 \,\cos(\omega(\tau_1 - \tau_2))\right] \quad \text{(A210)}$$

$$= \int_{-\infty}^{\infty} d\tau \,\exp\left[+i(\Omega - \Omega_{\rm peak})\tau - \frac{\Delta^2 k_BT}{\hbar^2}(1 - \cos(\omega\tau))\right]. \quad \text{(A211)}$$

Now we notice that the term $\exp[-(\Delta^2 k_BT/\hbar^2)\cos(\omega\tau)]$ is periodic, and thus has a discrete Fourier expansion; the only frequencies which appear are integer multiples of the vibrational frequency ω. As a result,

$$\sigma(\Omega) = \sum_n A_n \delta(\Omega - \Omega_{\rm peak} - n\omega). \quad \text{(A212)}$$

Thus, in this limit of a single undamped mode, the absorption spectrum *does* consist of a set of sharp lines, spaced by the vibrational quanta. In order to recover the semi–classical picture, these resonances must be washed out by a combination of multiple modes (so that the discrete absorption lines become a dense forest) and some dissipation corresponding to a lifetime or dephasing of each individual mode.

Problem 176: Washing out resonances. Suppose that we have just a single mode, but this mode is damped so that

$$\langle q(t)q(t')\rangle = \frac{k_BT}{\omega^2}\cos[\omega(t-t')]\exp\left[-\gamma|t-t'|\right]. \quad \text{(A213)}$$

If $\gamma \ll \omega$, the integral in Eq (A205) which defines the absorption cross–section is almost the integral of a period function. Thus there will be multiple saddle points, the first (the one we have considered in our semi–classical approximation) being close to $\tau = 0$, and all the others close to $\tau = 2\pi n/\omega$ for integer n. Carry out this expansion, and analyze your results. Can you see how, as $\gamma \to 0$, this sum over saddle points gives back the discrete spectral lines? At large γ, what enforces the smooth dependence of the cross–section on Ω? How big does γ need to be in order that we wouldn't see much hint of the vibrational resonances in the absorption spectrum? Is it possible that the vibrations are weakly damped ($\gamma \ll \omega$), but there are no visible resonances in the absorption spectrum?

Say something about the quantum treatment of the coordinate $q(t)$, and the zero–phonon lines. Maybe a word about the relation to the Moössbauer effect?

I think there is still more to say here. Notice that to make things consistent we need a quantum mechanical treatment of damping, which was a big puzzle some time back. This is also related to discussions of decoherence in more modern times. At the very least we need pointers to references. One could also note that what enters these computations are certain correlation functions of the "relevant" coordinates, and so if these correlation functions are damped (however this happens!) all will be well. Still ... an opportunity to teach some physics shouldn't be missed.

Need refs to standard text on molecular spectra; maybe old refs to solid state problem of electron–phonon couplings in impurity spectra. The idea that coupling to a bath of oscillators could describe dissipation in quantum mechanics goes back, at least, to Feynman & Vernon (1963). These ideas were revitalized by Caldeira & Leggett (1981, 1983), who were especially interested in the impact of dissipation on quantum tunneling.

Caldeira & Leggett 1981: Influence of dissipation on quantum tunnelling in macroscopic systems. AO Caldeira & AJ Leggett, *Phys Rev Lett* **46,** 211–214 (1981).

Caldeira & Leggett 1983: Quantum tunnelling in a dissipative system. AO Caldeira & AJ Leggett, *Ann Phys (NY)* **149,** 374–456 (1983).

Feynman & Vernon 1963: The theory of a general quantum system interacting with a linear dissipative system. RP Feynman & FL Vernon Jr, *Ann Phys (NY)* **24,** 118–173 (1963).

---

4. Cooperativity

[Be sure to talk about the specific case of hemoglobin, so we can point from Section II.A.]

To understand the statistical mechanics of cooperative interactions in the binding of multiple ligands, it is useful to start at the beginning, with the binding of a single ligand, especially since many physics students don't have much experience with problems that get categorized as "chemistry." Suppose that we have a receptor molecule R to which some smaller ligand molecule L can bind. For simplicity let there just be the two states, R with its binding site empty, and RL with the binding site filled by an L molecule, and let us assume that every binding event is independent, so the different receptor molecules don't interact. To study the dynamics of this system we keep track of the number of receptors in the state R and the number in state RL; these numbers, n_R and n_{RL}, respectively, must add up to give the total number of receptors, N.

The rate at which empty sites get filled ($R \rightarrow RL$) must be proportional to the number of empty sites and to the concentration c of the ligand. The rate at which filled sites become empty should just be proportional to the number of filled sites. Thus

$$\frac{dn_{RL}}{dt} = k_+ c n_R - k_- n_{RL}, \tag{A214}$$

where k_+ is the rate constant for binding and k_- is the rate constant for unbinding; note that these have different units. Since $n_R + n_{RL} = N$, this becomes

$$\frac{dn_{RL}}{dt} = k_+ cN - (k_- + k_+ c) n_{RL}. \tag{A215}$$

The equilibrium state is reached when

$$n_{RL} = N \frac{k_+ c}{k_- + k_+ c}. \tag{A216}$$

The fraction n_{RL}/N can also be interpreted microscopically as the probability that one receptor will be the state RL,

$$P_{RL} = \frac{k_+ c}{k_- + k_+ c} = \frac{c}{K + c}, \tag{A217}$$

where the equilibrium constant (or "dissociation constant") $K = k_-/k_+$.

From statistical mechanics, if we have a molecule that can be in two states, we should calculate the probability of being in these states by knowing the energy of each state and using the Boltzmann distribution. Importantly, what we mean by "state," especially when discussing large molecules, often is a large group of microscopic configurations. Thus saying that there are two states R and RL really means that we can partition the phase space of the system into two regions, and these regions are what we label as R and RL. Then, as should be familiar, what matters is not the energy of each state but the free energy. The free energy of the state R has one component from the receptor molecule itself, F_R, plus a component from the ligand molecules in solution. In the transition $R \rightarrow RL$, the free energy of the receptor changes to F_{RL}, *and* the free energy of the solution changes because one molecule of the ligand is removed. The change in free energy when we add one molecule to the solution defines the chemical potential $\mu(c)$. Thus, up to an arbitrary zero of energy, we can consider the free energy of the two states to be F_R and $F_{RL} - \mu(c)$. Then the probability of being in the state RL is given by the Boltzmann distirbution,

$$P_{RL} = \frac{1}{Z}\exp\left(-\frac{F_{RL} - \mu(c)}{k_BT}\right), \qquad \text{(A218)}$$

where the partition function Z is given by the sum of the Boltzmann factors over both available states,

$$Z = \exp\left(-\frac{F_R}{k_BT}\right) + \exp\left(-\frac{F_{RL} - \mu(c)}{k_BT}\right). \qquad \text{(A219)}$$

Putting the terms together, we have

$$P_{RL} = \frac{\exp\left[-(F_{RL} - \mu(c))/k_BT\right]}{\exp\left[-F_R/k_BT\right] + \exp\left[-(F_{RL} - \mu(c))/k_BT\right]} \qquad \text{(A220)}$$

$$= \frac{e^{\mu(c))/k_BT}}{\exp\left[-(F_R - F_{RL})/k_BT\right] + e^{\mu(c))/k_BT}}. \qquad \text{(A221)}$$

Notice that the only place where the ligand concentration appears is in the chemical potential $\mu(c)$. In order for this result to be consistent with the result from analysis of the kinetics in Eq (A217), we must have $e^{\mu(c))/k_BT} \propto c$, and you may recall that when concentrations are low—as in ideal gases, and also ideal solutions—it is a standard result that

$$\mu(c) = k_BT\ln(c/c_0), \qquad \text{(A222)}$$

where c_0 is some reference concentration. Then we can also identify the equilibrium constant as

$$K = c_0\exp\left(-\frac{F_{\text{bind}}}{k_BT}\right), \qquad \text{(A223)}$$

where $F_{\text{bind}} = F_R - F_{RL}$ is the change in free energy when the ligand binds to the receptor.

Now suppose we have a receptor to which two ligands can bind. There are now four states, which we can think of as 00, 10, 01, and 11. If the each binding event is identical and independent, then the free energies of these states are

$$F_{00} = F_R \qquad \text{(A224)}$$

$$F_{01} = F_{10} = F_R - F_{\text{bind}} - \mu(c) \qquad \text{(A225)}$$

$$F_{11} = F_R - 2F_{\text{bind}} - 2\mu(c). \qquad \text{(A226)}$$

If we calculate, for example, the probability that both binding sites are occupied—i.e., that the molecule is in the state 11—we have

$$P_{11} = \frac{1}{Z}e^{-F_{11}/k_BT} \qquad \text{(A227)}$$

$$= \frac{\exp\left[-\frac{F_R - 2F_{\text{bind}} - 2\mu(c)}{k_BT}\right]}{\exp\left[-\frac{F_R}{k_BT}\right] + 2\exp\left[-\frac{F_R - F_{\text{bind}} - \mu(c)}{k_BT}\right] + \exp\left[-\frac{F_R - 2F_{\text{bind}} - 2\mu(c)}{k_BT}\right]} \qquad \text{(A228)}$$

$$= \frac{(c/K)^2}{1 + 2(c/K) + (c/K)^2} = \left(\frac{c}{c+K}\right)^2. \qquad \text{(A229)}$$

Thus, the probability of both sites being occupied is just the square of the probability that a single binding site will be occupied, as in Eq (A217). This makes sense, because we assumed that binding to the two sites were independent events.

Problem 177: Counting bound molecules. Rather than counting the fraction of molecules in the doubly bound state, count the number of ligands bound. Show that this is just $2 \times c/(c+K)$, and explain why.

In fact, in many cases we see that binding of multiple ligands to a protein molecule are not independent events. As a start, let's suppose that we again have two binding sites, but the doubly bound state is stabilized (for as yet unspecified reasons) by an extra energy Δ. Then if we calculate the fraction of binding sites occupied, we have

$$f = \frac{1}{2}\left[P_{01} + P_{10} + 2P_{11}\right] \tag{A230}$$

$$= \frac{1}{2}\,\frac{2\exp\left[-\frac{F_R - F_{\rm bind} - \mu(c)}{k_BT}\right] + 2\exp\left[-\frac{F_R - 2F_{\rm bind} - 2\mu(c) - \Delta}{k_BT}\right]}{\exp\left[-\frac{F_R}{k_BT}\right] + 2\exp\left[-\frac{F_R - F_{\rm bind} - \mu(c)}{k_BT}\right] + \exp\left[-\frac{F_R - 2F_{\rm bind} - 2\mu(c) - \Delta}{k_BT}\right]} \tag{A231}$$

$$= \frac{c/K + J(c/K)^2}{1 + 2(c/K) + J(c/K)^2}, \tag{A232}$$

where $J = \exp(\Delta/k_BT)$. Results are shown in Fig 168. We see that, as the interaction energy increases, the binding sites can be occupied at lower concentration, but more importantly the steepness of the "switch" from empty to full sites is more abrupt. This abruptness is the signature of cooperativity.

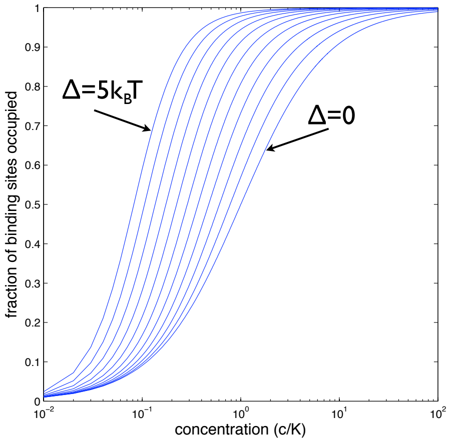

FIG. 168 Cooperative binding, with two binding sites that interact. Lines show the predicted fraction of binding sites vs. concentration, for different values of the interaction energy Δ; from Eq (A232).

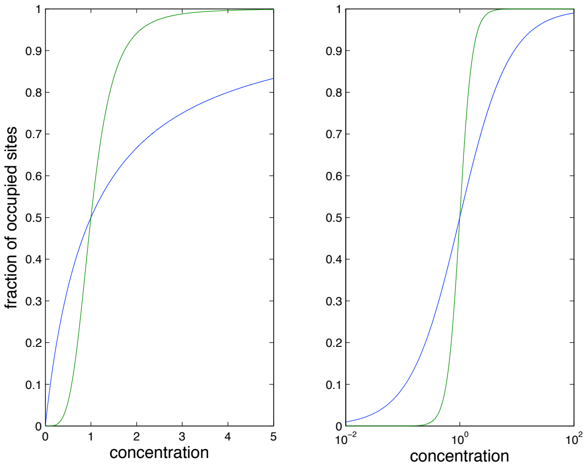

FIG. 169 Cooperative binding, in the Hill model. Blue lines show the predicted fraction of binding sites vs. concentration when binding to each site in independent. Green lines show the case of cooperative binding to four sites, as described by the Hill model in Eq (A233), with $n = 4$. At left, a linear concentration scale; at right, a logarithmic scale.

The classic example is the oxygen binding protein hemoglobin in our blood. We now know that hemoglobin has four protein subunits, each of which has an iron atom which can bind one oxygen molecule. [Might be good to show some figures from Hb!] As Hill recognized in the early part of the twentieth century, the fraction of sites with bound oxygen behaves more nearly as if all four molecules had to bind together, so that

$$f = \frac{c^n}{c^n + K^n}, \tag{A233}$$

with $n = 4$; this is still called a "Hill function" in many contexts. As shown in Fig 169, the binding is now sigmoidal, or more nearly switch like at larger n. Because the natural quantity in statistical mechanics is the chemical potential and not the concentration, things look simpler on a logarithmic concentration axis. Cooperative binding corresponds to a steeper slope on these logarithmic plots. This is clear for the Hill function, where we can see that

$$\frac{dF}{dc} = \frac{n}{c}F(1-F), \tag{A234}$$

and hence

$$\left.\frac{dF}{d\ln c}\right|_{F=1/2} = \frac{n}{4}, \tag{A235}$$

so the slope is a direct measure of the number of molecules forced to bind simultaneously. Few real systems are described exactly by the Hill model, but it's a good approximation. We should also appreciate the power of Hill's intuition, in seeing the connection of the sigmoidal binding curves to the number of protein subunits even before much was known about these molecules.

The Hill model suggests that there is some direct interaction between binding events that causes all of the ligands to bind (or not to bind) simultaneously, which we can think of as a limiting case of the model above, with $\Delta \to \infty$. In some cases, including hemoglobin, there is little evidence for such a direct interaction. An alternative is to imagine that the whole system can be in two states. In the case of hemoglobin these came to be called 'relaxed' (R) and 'tense' (T), but in other systems there natural choices; for example, in the case of the ion channels in rod cells that open in response to binding of cGMP, the two states might simply be the open and closed states of the channel, as in Fig 170. To continue with this example, the channel can bind one, two or three molecules of cGMP. If all the binding sites are empty, the free energies of the two states are $F_{\rm open}$ and $F_{\rm closed}$. Given that the channel is closed, the binding of a single cGMP molecule lowers the energy by an amount $F^{\rm bind}_{\rm closed}$, but in addition this takes one molecule out of the solution and hence the free energy of the system also goes down by μ, the chemical potential. So the total free energy of the state with the channel closed and one molecule bound is

$$\begin{aligned} F_{\rm closed}(1) &= F_{\rm closed} - F^{\rm bind}_{\rm closed} - \mu \qquad \text{(A236)} \\ &= F_{\rm closed} - F^{\rm bind}_{\rm closed} - k_BT\ln(c/c_0)\text{(A237)} \\ &= F_{\rm closed} - k_BT\ln\left(\frac{c}{K_{\rm closed}}\right), \qquad \text{(A238)} \end{aligned}$$

and similarly for the open state,

$$F_{\rm open}(1) = F_{\rm open} - k_BT\ln\left(\frac{c}{K_{\rm open}}\right). \tag{A239}$$

The important point is that the binding energies to the open and closed states are different. By detailed balance, this means that, as the cGMP molecules bind, they will shift the equilibrium between open and closed. The two state model was proposed by Monod, Wyman and Changeaux. They made the simplifying assumption that the only source of cooperativity among the binding events was this shifting of equilibria, so that if the target protein is in one state, each binding event remains independent, and then the free energies work out as in Fig 170.

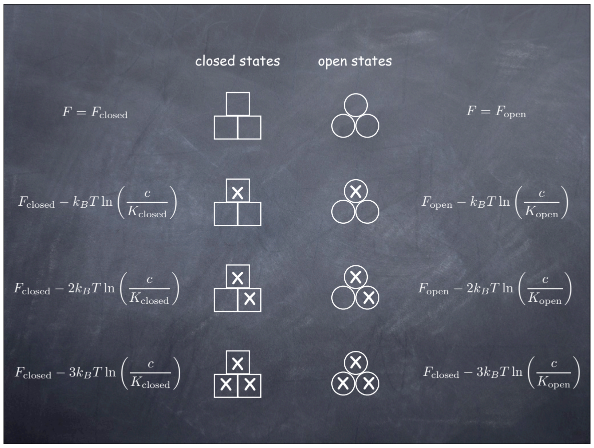
closed states
open states
$F = F_{\rm closed}$
$F = F_{\rm open}$
$F_{\rm closed} - k_BT\ln\left(\frac{c}{K_{\rm closed}}\right)$
$F_{\rm open} - k_BT\ln\left(\frac{c}{K_{\rm open}}\right)$
$F_{\rm closed} - 2k_BT\ln\left(\frac{c}{K_{\rm closed}}\right)$
$F_{\rm open} - 2k_BT\ln\left(\frac{c}{K_{\rm open}}\right)$
$F_{\rm closed} - 3k_BT\ln\left(\frac{c}{K_{\rm closed}}\right)$
$F_{\rm closed} - 3k_BT\ln\left(\frac{c}{K_{\rm open}}\right)$

FIG. 170 A model for binding of cGMP to the channels in rod cells. Cooperativity arises not from direct interactions among the cGMP molecules but rather because binding of each molecule contributes to stabilizing a different structure of the channel protein. In this case the two structures are just the open and closed states.

Problem 178: Cooperativity in the MWC model. Show that the model in Fig 170 is equivalent to the statement that the free energy difference between open and closed states has a term proportional to the number of cGMP molecules bound. What is this proportionality constant in terms of the other parameters? Can you explain the connection between these two points of view on the model?

It's a useful exercise to work out the statistical mechanics of the MWC model. The partition function has two classes of terms, coming from the two states of the protein. In each state, we have to sum over the occupied and unoccupied states of each binding site, but this is relatively easy because the sites are independent. In the notation of Fig 170, we find

$$\begin{aligned} Z &= Z_{\rm open} + Z_{\rm closed} \qquad \text{(A240)} \\ Z_{\rm open} &= \exp\left(-\frac{F_{\rm open}}{k_BT}\right)\left(1+\frac{c}{K_{\rm open}}\right)^n \qquad \text{(A241)} \\ Z_{\rm closed} &= \exp\left(-\frac{F_{\rm closed}}{k_BT}\right)\left(1+\frac{c}{K_{\rm closed}}\right)^n ,\text{(A242)} \end{aligned}$$

where in the case of the cGMP–gated channels, $n = 3$. The probability of being in the open state is then

$$\begin{aligned} P_{\rm open} &= \frac{Z_{\rm open}}{Z_{\rm open}+Z_{\rm closed}} \qquad \text{(A243)} \\ &= \frac{(1+c/K_{\rm open})^n}{(1+c/K_{\rm open})^n + L\,(1+c/K_{\rm closed})^n},\text{(A244)} \end{aligned}$$

where $k_BT\ln L = F_{\rm open} - F_{\rm closed}$ is the free energy difference between open and closed states in absence of ligand binding. In the limit that binding is much stronger to the open state, $K_{\rm open} \ll K_{\rm closed}$, this simplifies,

$$\begin{aligned} P_{\rm open} &= \frac{(1+c/K_{\rm open})^n}{L+(1+c/K_{\rm open})^n} \qquad \text{(A245)} \\ &= \frac{1}{1+\exp\left[\theta - n\ln(1+c/K_{\rm open})\right]}, \qquad \text{(A246)} \end{aligned}$$

where $\theta = \ln L$. This is similar to the Hill model, but a little different in detail. Distinguishing the models from the equilibrium data alone is difficult, but clearly the MWC model predicts that binding has an extra kinetic step in which the protein makes the transition between its two states; if we are lucky we can "catch" the system after the first ligand molecules have bound but before this change in protein structure. Indeed, such experiments were critical in understanding the mechanism of cooperativity in hemoglobin.

Problem 179: Details of the MWC model. Fill in the steps to Eqs (A240–A242). Then, compare the Hill model with MWC. Show that for $c \gg K_{\rm open}$, Eq (A246) reduces to Eq (A233). What about at $c \ll K_{\rm open}$? The MWC model, even in the limit $K_{\rm open} \ll K_{\rm closed}$, has one more parameter than the Hill model; what does this freedom mean for the class of functions that the MWC model can realize?

In many systems, it is not just a single class of ligands that binds. For hemoglobin itself, changes in pH, which presumably result in binding and unbinding of protons, change the way in which oxygen binds.[95] For enzymes—proteins that catalyze a chemical reaction—it is not just the substrate which binds and is chemically altered, but other molecules bind as well and alter the activity of the enzyme. It is important that these 'other molecules' are binding at other sites, not directly interfering with substrate binding in enzymes or oxygen binding in hemoglobin. From the Greek for "other site," these effects are called "allosteric," and the MWC model gives a framework for a much more general view of allostery. In this view, *all* binding events are independent, but with binding energies that depend on the overall state of the target protein. In this way, all binding events can shift the R/T equilibrium.

Maybe another problem? There should be a figure with data! Tell the story about Perutz? Need to flesh out the text to match references. Put something about protein/DNA interactions here?

The classic paper on "Hill functions" for cooperative binding is Hill (1910). There is some suggestion that Hill might have been the first to derive the simpler description of independent binding, often called the "Langmuir" isotherm; for this and more related history as seen through the lens of drug–receptor interactions, see Colquhoun (2006). The MWC model is due to Monod et al (1965), and a contemporary, competing model is due to Koshland et al (1966). Late in his life, Perutz (1990) provided some perspective on his long adventure with hemoglobin. A key step in understanding was to show, convincingly, that there really is no direct interaction between the binding sites, and the cooperativity was mediated entirely by the shifting equilibrium between the R and T states (Shulman et al 1975). The MWC model leaves open the question of where the energy for cooperativity is stored in the molecule; for a hypothesis very much ahead of its time, see Hopfield (1973).

Colquhoun 2006: The quantitative analysis of drug–receptor interactions: A short history. D Colquhoun, *Trends Pharm Sci* **27,** 149–157 (2006).

Hill 1910: The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. AV Hill, *J Physiol (Lond)* **40,** Suppl iv–vii (1910).

Hopfield 1973: Relation between structure, co-operativity and spectra in a model of hemoglobin action. JJ Hopfield, *J Mol Biol* **77,** 207–222 (1973).

Koshland et al 1966: Comparison of experimental binding data and theoretical models in proteins containing subunits. DE Koshland Jr, G Némethy & D Filmer, *Biochemistry* **5,** 365–385 (1966).

Monod et al 1965: On the nature of allosteric transitions: A plausible model. J Monod, J Wyman & JP Changeux, *J Mol Biol* **12,** 88–118 (1965).

Perutz 1990: *Mechanisms of Cooperativity and Allosteric Regulation in Proteins.* MF Perutz (Cambridge University Press, Cambridge, 1990).

Shulman et al 1975: Allosteric interpretation of haemoglobin properties. RG Shulman, JJ Hopfield & S Ogawa, *Q Rev Biophys* **8,** 325–420 (1975).

The idea that interesting things happen in "cooperative" events in larger and larger collections of interacting subunits probably occurred to many people who thought in terms of statistical mechanics (Thompson 1972), and it seems to get rediscovered periodically (Bray et al 1998, Duke & Bray 1999, Duke et al 2001). Statistical mechanics models for interactions among binding events also play a key role in thinking about protein/DNA interactions and the regulation of gene expression (Bintu et al 2005a,b, Kinney et al 2010).

Bintu et al 2005a: Transcriptional regulation by the numbers: models. L Bintu, NE Buchler, HG Garcia, U Gerland, T Hwa, J Kondev & R Phillips, *Curr Opin Gene Dev* **15,** 116–124 (2005).

Bintu et al 2005b: Transcriptional regulation by the numbers: applications. L Bintu, NE Buchler, HG Garcia, U Gerland, T Hwa, J Kondev, T Kuhlman & R Phillips, *Curr Opin Gene Dev* **15,** 125–136 (2005).

Bray et al 1998: Receptor clustering as a cellular mechanism to control sensitivity. D Bray MD Levin & CJ Morton–Firth, *Nature* **393,** 85–88 (1998).

Duke & Bray 1999: Heightened sensitivity of a lattice of membrane receptors. TAJ Duke & D Bray, *Proc Nat'l Acad Sci (USA)* **96,** 10104–10108 (1999).

Duke et al 2001: Conformational spread in a ring of proteins: A stochastic view of allostery. TAJ Duke, N Le Novère & D Bray, *J Mol Biol* **308,** 541–553 (2001).

Kinney et al 2010: Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. JB Kinney, A Murugan, CG Callan Jr & EC Cox, *Proc Nat'l Acad Sci (USA)* **107,** 9158–9163 (2010).

Thompson 1971: *Mathematical Statistical Mechanics* CJ Thompson (Macmillan, New York, 1972).

[95] This is the "Bohr effect," after Christian Bohr, Niels' father.

5. X–ray diffraction and biomolecular structure

The first detailed experimental information about the structure of biological molecules came from X–ray diffraction measurements. We recall that if a particle scatters from a sample, shifting its energy by $\hbar\omega$ and its momentum by $\hbar\vec{\mathbf{q}}$, then the amplitude for this scattering event must be proportional to the $(\vec{\mathbf{q}},\omega)$ spatiotemporal Fourier component of the relevant density in the sample. For an electromagnetic wave what matters is (roughly) the charge density. Thus, the cross–section for elastic $(\omega = 0)$ scattering is

$$\sigma(\vec{\mathbf{q}}) \propto \left| \int d^3x\, e^{i\vec{\mathbf{q}}\cdot\vec{\mathbf{x}}} \rho(\vec{\mathbf{x}}) \right|^2 . \tag{A247}$$

It is useful to have in mind the geometry [ref to a Fig!]. If the X–ray photons approach the sample collimated along the $\hat{x}$ axis, they have an initial wavevector $\vec{\mathbf{k}}_0 = k\hat{x}$, where as usual $k = 2\pi/\lambda$, with λ the wavelength. If they emerge with a final wavevector $\vec{\mathbf{k}}_{\rm f}$ at an angle θ relative to the $\hat{x}$ axis, then $\vec{\mathbf{q}} \equiv \vec{\mathbf{k}}_{\rm f} - \vec{\mathbf{k}}_0$, and the magnitude of the scattering vector (or, up to a factor $\hbar$, momentum transfer) is

$$\begin{aligned} |\vec{\mathbf{q}}| &= |\vec{\mathbf{k}}_{\rm f} - \vec{\mathbf{k}}_0| && \text{(A248)}\\ &= \sqrt{|\vec{\mathbf{k}}_{\rm f} - \vec{\mathbf{k}}_0|^2} && \text{(A249)}\\ &= \sqrt{|\vec{\mathbf{k}}_{\rm f}|^2 - 2\vec{\mathbf{k}}_{\rm f}\cdot\vec{\mathbf{k}}_0 + |\vec{\mathbf{k}}_0|^2} && \text{(A250)}\\ &= \sqrt{k^2 - 2k^2\cos\theta + k^2} && \text{(A251)}\\ &= \sqrt{2k^2(1-\cos\theta)} = 2k\sin(\theta/2). && \text{(A252)} \end{aligned}$$

Thus scattering by a small angle corresponds to a small momentum transfer. The classic results about X–ray diffraction concern the case where the density profile is periodic, as in a crystal. If the periodicity corresponds to displacement by d (let's think along one dimension, for the moment), then the density can be expressed as a discrete Fourier series, which means [from Eq (A247)] that $\sigma(\vec{\mathbf{q}})$ will have delta functions at $|\vec{\mathbf{q}}| = 2\pi n/d$, with n an integer. Combining this with Eq (A252), we find the angles which satisfy the "Bragg condition,"

$$2\pi n/d = (4\pi/\lambda)\sin(\theta/2) \Rightarrow \sin(\theta/2) = n\lambda/2d. \tag{A253}$$

[I think this is a bit off the usual way of stating the condition (2's in the wrong places); check!]

The first great triumph of X–ray diffraction in elucidating the structure of biological molecules came with the structure of DNA. This is an often told, and often distorted, piece of scientific history. Watson and Crick predicted the structure of DNA by arguing that a few key facts about the molecule, when combined with the rules of chemical bonding, where enough to suggest an interesting structure that would have consequences for the mechanisms of genetic inheritance. It was known that DNA was composed of four different kinds of nucleotide bases: adenine (A), thymine (A), guanine (G) and cytosine (C). Importantly, Chargaff had surveyed the DNA of many organisms and shown that while the ratios of A to G, for example, vary enormously, the ratios A/T and C/G do not. Watson and Crick realized that the molecular structures of the bases are such that A and T can form favorable hydrogen bonds, as can C and G; further, the resulting hydrogen bonded base pairs are the same size, and thus could fit comfortably into a long polymer, as shown in Fig 171. Piling on top of one another, the base pairs would also experience a favorable "stacking" interaction among the π–bonded electrons in their rings. Finally, if one looks carefully at all the bond angles where the planar bases connect to the sugars and phosphate backbone, each successive base pair must rotate relative to its neighbor, and although there is some flexibility the favored angle was predicted to be $2\pi/10$ radians, or 36°.

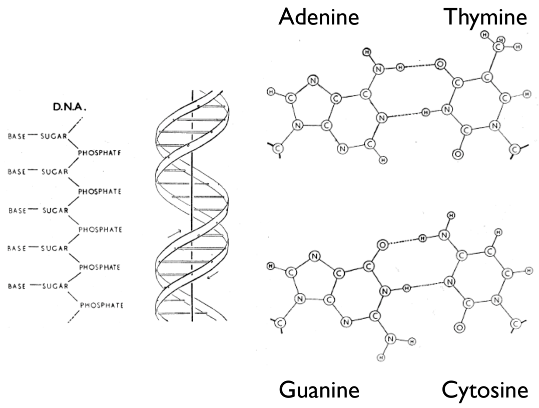

FIG. 171 The structure of DNA, from Watson and Crick (1953b). At left, the polymeric pattern of bases, sugars and phosphates, and the famous double helix. At right, the pairings A/T and G/C, illustrating the similar sizes of the correct pairs. Note that the donor/acceptor pattern of hydrogen bonds discriminates against the incorrect A/C and G/T pairings.

Quite independently of his collaboration with Watson, Crick has been interested in the structure of helical molecules, and in the X–ray diffraction patterns that they should produce. Thus, when Watson and Crick realized that the structure of DNA might be a helix, they were in a position to calculate what the diffraction patterns should look like, and thus compare with the data emerging from the work of Franklin, Wilkins and collaborators. So, let's look at the theory of diffraction from a helix.

It's best to describe a helix in cylindrical coordinates: z along the axis of the helix, r outward from its center, and an angle ϕ around the axis. Helical symmetry is the statement that translations along z are equivalent to rotations of the angle ϕ. Thus, a continuous helical structure would have the property that

$$\rho(z, r, \phi) = \rho(z + d, r, \phi + 2\pi d/\ell), \tag{A254}$$

for any displacement d, where ℓ is the displacement corresponding to a complete rotation. For a discrete helical structure, the same equation is true, but only for values of d that are integer multiples of a fundamental spacing d_0.

For the continuous helix, the dependence on the two variables z and ϕ really collapses to a dependence on one combined variable,

$$\rho(z, r, \phi) = g(r, \phi - 2\pi z/\ell). \tag{A255}$$

We know that any function of angle can be expanded as a discrete Fourier series,

$$f(\phi) = \sum_{n=-\infty}^{\infty} \tilde{f}_n e^{-in\phi}, \tag{A256}$$

so in this case we have

$$\rho(z, r, \phi) = \sum_{n=-\infty}^{\infty} \tilde{g}_n(r) e^{-in(\phi - 2\pi z/\ell)}. \tag{A257}$$

Our task is to compute

$$\int d^3x \, e^{i\vec{\mathbf{q}}\cdot\vec{\mathbf{x}}} \rho(\vec{\mathbf{x}}). \tag{A258}$$

In cylindrical coordinates, we can write $\vec{\mathbf{q}} = (q_z\hat{z}, \vec{\mathbf{q}}_\perp)$, so that $\vec{\mathbf{q}}\cdot\vec{\mathbf{x}} = q_z z + q_\perp r\cos\phi$, where we choose the origin of the angle ϕ to make things simple and $q_\perp = |\vec{\mathbf{q}}_\perp|$. Thus we have

$$e^{i\vec{\mathbf{q}}\cdot\vec{\mathbf{x}}} = e^{iq_z z} e^{iq_\perp r\cos\phi} \tag{A259}$$

$$= e^{iq_z z} \sum_{n=-\infty}^{\infty} J_n(q_\perp r) e^{in\phi}, \tag{A260}$$

where [check the conventions for the definition of the Bessel function!]

$$J_n(u) = \int_0^{2\pi} \frac{d\phi}{2\pi} e^{-in\phi} e^{iu\cos\phi} \tag{A261}$$

are Bessel functions. Putting Eq (A260) together with the consequences of helical symmetry in Eq (A257), we have

$$\int d^3x \, e^{i\vec{\mathbf{q}}\cdot\vec{\mathbf{x}}} \rho(\vec{\mathbf{x}}) = \int_{-\infty}^{\infty} dz \int_0^{\infty} dr r \int_0^{2\pi} d\phi e^{iq_z z} \sum_{n=-\infty}^{\infty} J_n(q_\perp r) e^{in\phi} \sum_{m=-\infty}^{\infty} \tilde{g}_m(r) e^{-im(\phi - 2\pi z/\ell)} \tag{A262}$$

$$= \sum_{n,m-\infty}^{\infty} \int_{-\infty}^{\infty} dz \, e^{iq_z z} e^{-i2\pi m z/\ell} \int_0^{\infty} dr r \, J_n(q_\perp r)\tilde{g}_m(r) \int_0^{2\pi} e^{in\phi} e^{-im\phi}. \tag{A263}$$

We see that the integral over ϕ forces $m = n$, and the integral over z generates delta functions at $q_z = 2\pi n/\ell$. Thus, for a continuous helix we expect that the X–ray scattering cross section will behave as

$$\sigma(q_z, q_\perp) \propto \sum_{n-\infty}^{\infty} \delta(q_z - 2\pi n/\ell) \left| \int_0^{\infty} dr r \, J_n(q_\perp r)\tilde{g}_n(r) \right|^2. \tag{A264}$$

In particular, if most of the density sits at a distance R from the center of the helix (which is not a bad approximation for DNA, since the phosphate groups have much more electron density than the rest of the molecule), then

$$\sigma(q_z, q_\perp) \sim \sum_{n-\infty}^{\infty} \delta(q_z - 2\pi n/\ell) \left| J_n(q_\perp R) \right|^2. \tag{A265}$$

Equation (A265) is telling us that diffraction from a helix generates a series of "layer lines" at $q_z = 2\pi n/\ell$, and from their spacing we should be able to read off the "pitch" of the helix, the distance ℓ along the $\hat{z}$ axis corresponding to a complete turn. Further, if we look along a single layer line, we should see an intensity varying as

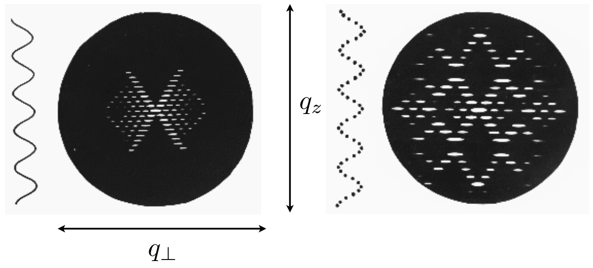

FIG. 172 Diffraction from continuous (left) and discrete (right) helices; Holmes (1998).

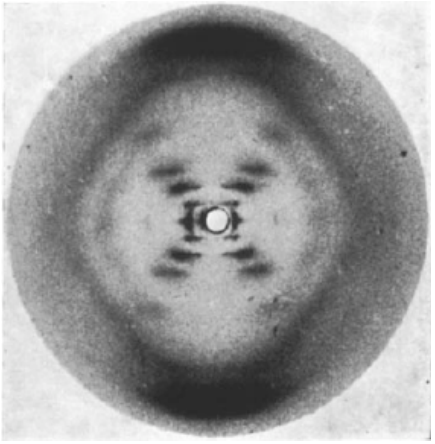

FIG. 173 The justly famous photograph 51, showing the diffraction from DNA molecules pulled into a fiber, from Franklin & Gosling (1953).

$\sim |J_n(q_\perp R)|^2$. What is important here about the Bessel functions is that for small $q_\perp$ we have $J_n(q_\perp R) \propto (q_\perp R)^n$, and the first peak of the $n^{\rm th}$ Bessel function occurs at a point roughly proportional to n. The resulting pattern is shown schematically in Fig 172.

Problem 180: Bessel functions. Verify the statements about Bessel functions made above, in enough detail to understand the diffraction patterns shown in Fig 172.

Let's see what happens when we move from the continuous to the discrete helix. To keep things simple, suppose that all the density indeed is concentrated at a distance R from the center of the helix, so that

$$\rho(\vec{\mathbf{x}}) = \frac{1}{R}\delta(r-R)\sum_n \delta(z - nd_0)\delta(\phi - n\phi_0), \qquad \text{(A266)}$$

where the rotation from one element to the next $\phi_0 = 2\pi d_0/\ell$; notice that we don't really require ℓ/d_0 to be an integer. Now we have

$$\int d^3x\, e^{i\vec{\mathbf{q}}\cdot\vec{\mathbf{x}}}\rho(\vec{\mathbf{x}}) = \int_{-\infty}^{\infty} dz \int_0^\infty drr \int_0^{2\pi} d\phi e^{iq_z z} \sum_{n=-\infty}^{\infty} J_n(q_\perp r)e^{in\phi}\frac{1}{R}\delta(r-R)\sum_{m=-\infty}^{\infty}\delta(z-md_0)\delta(\phi - m\phi_0) \qquad \text{(A267)}$$

$$= \sum_{n=-\infty}^{\infty} J_n(q_\perp R) \sum_{m=-\infty}^{\infty} \int_{-\infty}^{\infty} dz\, \delta(z-md_0)e^{iq_z z} \times \int_0^{2\pi} d\phi\, \delta(\phi - m\phi_0)e^{in\phi} \qquad \text{(A268)}$$

$$= \sum_{n=-\infty}^{\infty} J_n(q_\perp R) \sum_{m=-\infty}^{\infty} e^{im(n\phi_0 + q_z d_0)} \qquad \text{(A269)}$$

$$= \sum_{n=-\infty}^{\infty} J_n(q_\perp R) \sum_{m=-\infty}^{\infty} \delta(n\phi_0 + q_z d_0 - 2\pi m) \qquad \text{(A270)}$$

$$\propto \sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty} J_n(q_\perp R)\delta(q_z + 2\pi n/\ell - 2\pi m/d_0). \qquad \text{(A271)}$$

Thus the discrete helix involves a double sum of terms. If we set $m = 0$ we have the results for the continuous helix. But the sum over $m \neq 0$ causes the whole "X" pattern of the continuous helix to be repeated with centers at $(q_z = 2\pi m/d_0, q_\perp = 0)$; the line $q_\perp = 0$ is often called the meridian, and so the extra peaks centered on $(q_z = 2\pi m/d_0, q_\perp = 0)$ are called meridional refections. All of this is shown in Fig 172. Just as the spacing of the layer lines allows us to measure the helical pitch ℓ, the spacing of the meridional reflections allows us to measure the spacing d_0 between discrete elements along the helix.

At this point you know what Watson and Crick knew [maybe put in the precise dates of these events, from Watson's memoir]. They had a theory of what the structure should be, and almost certainly they had already realized the implications of this structure, as they remarked in their first paper "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material." They also knew that if the structure was as they had theorized, then the diffraction pattern should display a number of key signatures—the regularly spaced layer lines, the "X" arrangement of their intensities, and the meridional reflections—that would provide both qualitative and quantitative confirmation of the theory. Thus you should be able to imagine their excitement when they saw the clean X–ray diffraction pattern from hydrated DNA, the famous photograph 51 taken by Rosalind Franklin, Fig 173. As far as one could tell, the proposed structure was right.

Problem 181: Discrete helices, more generally. Show that most of what was said above can be generalized to an arbitrary discrete helix, without assuming that the density is concentrated at $r = R$. That is, use only the symmetry defined by Eq (A254) for $d = nd_0$.

Problem 182: Fibers vs. crystals. We have discussed the diffraction from a helix as if there were just one molecule, and we have not been very precise about the different between amplitudes and intensities. Show that if there are many helices, all with their $\hat{z}$ axes aligned but with random positions and orientations in the $\hat{x} - \hat{y}$ plane, then the diffraction intensity from the ensemble of molecules depends only on the structure of the individual helices, and that all directions for the vector $\vec{\mathbf{q}}_\perp$ are equivalent.

It is crucial to appreciate that, contrary to what is often said in textbooks, it was not possible to "determine" the structure of DNA by looking at diffraction patterns like those in Fig 173. On the other hand, if you thought you knew the structure, you could predict the diffraction pattern—in the regime where it could be measured—and see if you got things right. This difference between experiments that support a theory, or which find something that a theory tells us must exist, and experiments that "discover" something unexpected or genuinely unknown is an incredibly important distinction, often elided.

So much has been written about this moment in scientific history that it would be irresponsible not to pause and reflect. On the other hand, I am not a historian. So let me make make just a few observations. Most importantly, I think, the story of the DNA structure combines so many themes in our understanding of science and society (separately and together) that is has an almost mythical quality, and as with the ancient myths everyone can see something that connects to their own concerns. There is the enormous issue of gender in the scientific community, something for which we hardly even had a vocabulary until decades after the event. There are the personalities of all the individuals, both as they were in 1953 and as they developed in response to the world–changing discovery in which they participated. There is the tragedy of Franklin's early death. There is the competition between Cambridge and London, and the impact of an American interloper on these very British social structures. Finally, there are issues that are more purely about the science, such as the interaction between theory and experiment, physics and biology. We could wander in this part of history for a long time. I need to come back and see what is essential, and what can be skipped. For now, let's move on.

In order to actually **determine** the structure of a large molecule by X–ray diffraction, we need to form crystals of those molecules. Crystals of a protein are not like crystals of salt or even small molecules. They are quite soft, and contain quite a lot of water. The bonds between proteins, for example, in a crystal are much weaker than the bonds that hold each protein together. On the one hand this makes growing and handling the crystals quite difficult. On the other hand, it means that the internal structures of the protein in the crystal is more likely to be typical of its structure when free in solution.

We recall that being a crystal in three dimensions means that there are vectors $\vec{\mathbf{a}}$, $\vec{\mathbf{b}}$, and $\vec{\mathbf{c}}$ such that the density is the same if we translate by integer combinations of these vectors,

$$\rho(\mathbf{x}) = \rho(\mathbf{x} + n\vec{\mathbf{a}} + m\vec{\mathbf{b}} + k\vec{\mathbf{c}}). \tag{A272}$$

This means that the density can be expanded into a Fourier series,

$$\rho(\mathbf{x}) = \sum_{knm} \tilde{\rho}_{knm} \exp\left[i(k\vec{\mathbf{G}}_a + n\vec{\mathbf{G}}_b + m\vec{\mathbf{G}}_c)\cdot\vec{\mathbf{x}}\right], \tag{A273}$$

where the $\vec{\mathbf{G}}_{\rm i}$ are the "reciprocal lattice vectors." As a result, the X–ray scattering cross–section is a set of delta functions or "Bragg peaks,"

$$\sigma(\vec{\mathbf{q}}) \propto \sum_{knm} |\tilde{\rho}_{knm}|^2 \delta(\vec{\mathbf{q}} - k\vec{\mathbf{G}}_a - n\vec{\mathbf{G}}_b - m\vec{\mathbf{G}}_c). \tag{A274}$$

Problem 183: Details of diffraction. Fill in the details leading to Eq (**??**), including the relationship between the reciprocal lattice vectors $\vec{\mathbf{G}}_{\rm i}$ and the real lattice vectors $\vec{\mathbf{a}}$, $\vec{\mathbf{b}}$, and $\vec{\mathbf{c}}$.

Even if we can make a perfect measurement of $\sigma(\vec{\mathbf{q}})$, we only learn about the magnitudes of the Fourier coefficients, $|\tilde{\rho}_{knm}|^2$, and this isn't sufficient to reconstruct the density $\rho(\vec{\mathbf{x}})$. This is called the phase problem. For small structures it is not such a serious problem, since the constraint that $\rho(\vec{\mathbf{x}})$ has to built out of discrete atoms allows us to determine the positions of the atoms from the diffraction pattern. But for a protein, with thousands of atoms in each unit cell of the crystal, this is hopeless.

The phase problem was solved experimentally through the idea of "isomorphous replacement." Suppose that we could attach to the each molecule in the crystal one or more very heavy atoms, in well defined (but unknown) positions. If we can do this without disrupting the packing of the molecules into the crystal, then the positions of the Bragg peaks will not change, but their intensities will. If we can approximate the density profiles of the heavy atoms as delta functions (which should be right unless we look at very large $|\vec{\mathbf{q}}|$), then

$$|\tilde{\rho}_{knm}|^2 \rightarrow \left|\rho_{knm} + \sum_\mu Z_\mu e^{i\vec{\mathbf{q}}_{knm}\cdot\vec{\mathbf{x}}_\mu}\right|^2, \tag{A275}$$

where $\vec{\mathbf{q}}_{knm} = k\vec{\mathbf{G}}_a - n\vec{\mathbf{G}}_b - m\vec{\mathbf{G}}_c$, Z_μ is the charge of the $\mu^{\rm th}$ heavy atom and $\vec{\mathbf{x}}_\mu$ is its position. In the simple case of one added heavy atom, we can choose coordinates so that its position is at the origin, and then it should be clear that the change in intensity on adding the heavy atom is directly sensitive to the value of $\cos\phi_{knm}$, where ϕ_{knm} is the phase of the complex number ρ_{knm}. Thus, one needs at least two different examples of adding heavy atoms to determine the phases unambiguously.

Do we need to say more here? Show in detail how two replacements determines the phase? Give a problem? I honestly don't know if one has to rely on absolute measurements, as one might think naively from the equations ... check!! Say something about other approaches to the phase problem.

The density really consists of discrete blobs corresponding to atoms, and—if we can look at sufficiently high resolution—additional density in the bonds between atoms. For the moment let's think just about the atoms. Then the density has the form

$$\rho(\vec{\mathbf{x}}) \approx \sum_\mu f_\mu \delta(\vec{\mathbf{x}} - \vec{\mathbf{x}}_\mu), \tag{A276}$$

where $\vec{\mathbf{x}}_\mu$ is the position of the $\mu^{\rm th}$ atom and f_μ is an effective charge or scattering density associated with that atom. Thus the scattering cross–section behaves as

$$\sigma(\vec{\mathbf{q}}) \sim \sum_{\mu\nu} f_\mu f_\nu e^{i\vec{\mathbf{q}}\cdot(\vec{\mathbf{x}}_\mu - \vec{\mathbf{x}}_\nu)}. \tag{A277}$$

Importantly, the positions of atoms fluctuate. The time scale of these fluctuations typically is much shorter than the time scale of the experiment, so we will see an average,

$$\sigma(\vec{\mathbf{q}}) \sim \left\langle \sum_{\mu\nu} f_\mu f_\nu e^{i\vec{\mathbf{q}}\cdot(\vec{\mathbf{x}}_\mu - \vec{\mathbf{x}}_\nu)} \right\rangle. \tag{A278}$$

If we assume that the fluctuations in position are Gaussian around some mean, then

$$\begin{aligned} \sigma(\vec{\mathbf{q}}) &\sim \left\langle \sum_{\mu\nu} f_\mu f_\nu e^{i\vec{\mathbf{q}}\cdot(\vec{\mathbf{x}}_\mu - \vec{\mathbf{x}}_\nu)} \right\rangle \\ &\equiv \sum_{\mu\nu} f_\mu f_\nu \left\langle e^{i\vec{\mathbf{q}}\cdot\vec{\mathbf{r}}_{\mu\nu}} \right\rangle \qquad \text{(A279)} \\ &\sim \sum_{\mu\nu} f_\mu f_\nu e^{i\vec{\mathbf{q}}\cdot\vec{\mathbf{r}}_{\mu\nu}} e^{-\frac{1}{2}|\vec{\mathbf{q}}|^2\langle(\delta\vec{\mathbf{r}}_{\mu\nu})^2\rangle}, \qquad \text{(A280)} \end{aligned}$$

where $\vec{\mathbf{r}}_{\mu\nu} = \vec{\mathbf{x}}_\mu - \vec{\mathbf{x}}_\nu$, and for simplicity we assume that the fluctuations are isotropic. What we see is that the scattering intensity at $\vec{\mathbf{q}}$ is attenuated relative to what we expect from a fixed structure, by an amount $e^{-\frac{1}{2}|\vec{\mathbf{q}}|^2\langle(\delta\vec{\mathbf{r}}_{\mu\nu})^2\rangle}$. These are called the Debye–Waller factors. Thus, although X–ray diffraction is a static method, it is sensitive to dynamical fluctuations in structure, although it can't really distinguish between dynamics and static disorder in the crystal.

Need to come back and see what else needs to be said, given what we need in the main text. Is it worth talking about other methods, such as EM and NMR? The motifs of protein structure? ... not sure what we need or want.

You should read the classic trio of papers on DNA structure, which appeared one after the other in the April 25, 1953 issues of *Nature*: Watson & Crick (1953a), Wilkins et al (1953) and Franklin & Gosling (1953). The foundations of helical diffraction theory had been given just a year before by Cohcran et al (1952); a brief account is given by Holmes (1998). The astonishing realization that the structure of DNA implies a mechanism for the transmission of information from generation to generation was presented by Watson & Crick (1953b). It is especially interesting to read their account of the questions raised by their proposal, and to see how their brief list became the agenda for the emerging field of molecular biology over the next two decades. The rest is history, as the saying goes, so you should read at least one history book (Judson 1979).

Cochran et al 1952: The structure of synthetic polypeptides. I. The transform of atoms on a helix. W Cochran, FHC Crick & V Vand, *Acta Cryst* **5,** 581–586 (1952).

Franklin & Gosling 1953: Molecular configuration in sodium thymonucleate. RE Franklin & RG Gosling. *Nature* **171,** 740–741 (1953).

Holmes 1998: Fiber diffraction. KC Holmes, `http://www.mpimf-heidelberg.mpg.de/~holmes/fibre/branden.html` (1998).

Judson 1979: *The Eighth Day of Creation* HF Judson (Simon and Schuster, New York, 1979).

Watson & Crick 1953a: A structure for deoxyribose nucleic acid. JD Watson & FHC Crick, *Nature* **171,** 737–739 (1953).

Watson & Crick 1953b: Genetical implications of the structure of deoxyribonucleic acid. JD Watson & FHC Crick, *Nature* **171,** 964–967 (1953).

Wilkins et al 1953: Molecular structure of deoxypentose nucleic acids. MHF Wilkins, AR Stokes & HR Wilson, *Nature* **171,** 738–740 (1953).

Need classic refs about protein structure and crystallography; more if we do more.

---

6. Berg and Purcell, revisited

In the spirit of Berg and Purcell's original discussion, the simplest example of noise in a chemical system is just to consider the fluctuations in concentration as seen in a small volume. To treat this rigorously, let's remember that diffusion in and out of the volume keeps the system at equilibrium. Thus, fluctuations in the concentration should be just like Brownian motion or Johnson noise. What's a little odd is that while the strength of Johnson noise is proportional to the absolute temperature, our intuition about counting molecules and the $\sqrt{N}$ rule doesn't seem to have a place for T. So, let's see how this works.[96]

If we measure the current flowing across a resistor in thermal equilibrium at temperature T, we will find a noise in the current that has a spectral density $S_I = 2k_BT/R$, where R is the resistance. More generally, if we measure between two points in a circuit, and find a frequency dependent, complex impedance $\tilde{Z}(\omega)$, then the spectral density of current noise will be

$$S_I(\omega) = 2k_BT\mathrm{Re}\left[\frac{1}{\tilde{Z}(\omega)}\right], \tag{A281}$$

where Re denotes the real part. In a mechanical system it is more natural to talk about positions and forces instead of currents and voltages. Now if we measure the position and apply a force, we have a "mechanical response function" $\tilde{\alpha}(\omega)$ analogous to the (inverse) impedance,

$$\tilde{x}(\omega) = \tilde{\alpha}(\omega)\tilde{F}(\omega), \tag{A282}$$

where $\tilde{x}(\omega)$ is the Fourier component[97] of $x(t)$,

$$x(t) = \int_{-\infty}^{\infty}\frac{d\omega}{2\pi}e^{-i\omega t}\tilde{x}(\omega), \tag{A283}$$

and similarly for the force $\tilde{F}(\omega)$. The analog of Eq (A281) for Johnson noise is that the fluctuations in position x have a spectral density

$$S_x(\omega) = \frac{2k_BT}{\omega}\mathrm{Im}\left[\tilde{\alpha}(\omega)\right]. \tag{A284}$$

[It's possible that this Appendix should also contain a derivation of the FDT.]

Problem 184: Some details about noise spectra. You may remember the formula for Johnson noise as $S_I = 4k_BT/R$, rather than the factor of 2 given above. Also, there are a few obvious differences between Eqs (A281) and (A284). Be sure you understand all these differences. The key ingredients are that all our integrals run over positive and negative frequencies, and that while voltage is analogous to force, current is analogous to velocity, not position. Check carefully that all the details work out.

In any system at thermal equilibrium, if we apply a small force we can observe a proportionally small displacement, and this is described by a linear response function. In a mechanical system we have the function $\tilde{\alpha}(\omega)$, sometimes called a "complex compliance." In magnetic systems, the force is an applied magnetic field and the analog of position is the magnetization; the response function is called the susceptibility. Electrical systems are a bit odd because we usually discuss the current response to voltage, but we can also think about charge movements (see problem above). In all these cases, once we know the linear response function we can predict the spectral density of fluctuations in the relevant position–like variable using Eq (A284). This is called the fluctuation dissipation theorem. [Should there be an appendix with more details, and a proof? Advice welcome.]

Problem 185: Recovering equipartition. If we go to zero frequency, we have $\tilde{x} = \tilde{\alpha}(0)\tilde{F}$, but this means that $\tilde{\alpha}(0) = 1/\kappa$, where κ is the stiffness of the system. We know from the equipartition theorem that the variance in position must be related to the stiffness,

$$\frac{1}{2}\kappa\langle x^2\rangle = \frac{1}{2}k_BT. \tag{A285}$$

But we can also write the variance in position as an integral over the spectral density,

$$\langle x^2\rangle = \int_{-\infty}^{\infty}\frac{d\omega}{2\pi}S_x(\omega). \tag{A286}$$

For these equations to be consistent, we must have

$$2\int_{-\infty}^{\infty}\frac{d\omega}{2\pi}\frac{1}{\omega}\mathrm{Im}\left[\tilde{\alpha}(\omega)\right] = \tilde{\alpha}(0), \tag{A287}$$

which looks quite remarkable.

(a.) The frequency domain Eq (A282) is equivalent to

$$x(t) = \int_{-\infty}^{\infty}d\tau\alpha(\tau)F(t-\tau), \tag{A288}$$

where

$$\alpha(\tau) = \int_{-\infty}^{\infty}\frac{d\omega}{2\pi}e^{-i\omega\tau}\tilde{\alpha}(\omega). \tag{A289}$$

Causality means that $\alpha(\tau < 0) = 0$. What does this imply about the analytic properties of the $\tilde{\alpha}(\omega)$ in the complex ω plane?

(b.) Use your result in (a.) to verify Eq (A287).

Position and force, magnetization and magnetic field, charge and voltage; all of these are "thermodynamically conjugate" pairs of variables. More precisely, if we consider an ensemble in which the force is held fixed, then the derivative of the free energy with respect to the force

[96] In what follows I make free use of the concepts of correlation functions, power spectra, and all that. See Appendix A.2 for a review of these ideas.

[97] Here, more than in other sections, our conventions in defining the Fourier transform are important. Be careful about the sign of i in the exponential!

is the mean position, and conversely. The fluctuation dissipation theorem always refers to these pairs of variables. So, to describe fluctuations in chemical systems, we need to know the "force" that is conjugate to the concentration (or the number of molecules), and this is the chemical potential μ. To compute the response of the concentration to changes in chemical potential, we consider the diffusion equation for the concentration $c(\vec{x},t)$ in the presence of a varying chemical potential $\mu(\vec{x},t)$,

$$\frac{\partial c(\vec{x},t)}{\partial t} = D\nabla\cdot\left[\nabla c(\vec{x},t) - \frac{\nabla\mu(\vec{x})}{k_BT}c(\vec{x},t)\right]. \qquad \text{(A290)}$$

Problem 186: Connecting back. Explain how Eq (A290) relates to the equation for diffusion in the presence of an external potential, Eq (240). Be sure you understand the signs.

Linearizing Equation (A290) around a mean concentration $\bar{c}$, we have

$$\frac{\partial c(\vec{x},t)}{\partial t} = D\nabla^2 c(\vec{x},t) - \frac{D\bar{c}}{k_BT}\nabla^2\mu(\vec{x}). \qquad \text{(A291)}$$

We can solve by Fourier transforming in both space and time,

$$c(\vec{x},t) = \int\frac{d^3k}{(2\pi)^3}\int\frac{d\omega}{2\pi}e^{-i\omega t}e^{+i\vec{k}\cdot\vec{x}}\tilde{c}(\vec{k},\omega), \qquad \text{(A292)}$$

to find

$$\tilde{c}(\vec{k},\omega) = \frac{D\bar{c}}{k_BT}\frac{k^2}{-i\omega + Dk^2}\tilde{\mu}(\vec{k},\omega). \qquad \text{(A293)}$$

Thus there is a $\vec{k}$–dependent response function,

$$\tilde{\alpha}(\vec{k},\omega) = \frac{D\bar{c}}{k_BT}\frac{k^2}{-i\omega + Dk^2} \qquad \text{(A294)}$$

from which we can use the fluctuation dissipation theorem to calculate the spatiotemporal power spectrum of concentration fluctuations,

$$S_c(\vec{k},\omega) = \frac{2k_BT}{\omega}\mathrm{Im}\left[\tilde{\alpha}(\vec{k},\omega)\right] = 2\bar{c}\frac{Dk^2}{\omega^2 + (Dk^2)^2}. \qquad \text{(A295)}$$

Notice that the factors of k_BT cancel: the fluctuations are proportional to the temperature, but the response function—the susceptibility of the concentration to changes in chemical potential—is inversely proportional to the temperature.

How does the result in Eq (A295) relate to our intuition about the $\sqrt{N}$ rule? Let's think about measuring the average concentration in a small volume, which corresponds to the heuristic calculation by Berg and Purcell. To do this we construct a variable

$$C(t) = \int d^3x W(\vec{x})c(\vec{x},t), \qquad \text{(A296)}$$

where the weighting function $W(\vec{x})$ is $1/V$ inside a volume V, and zero outside. Then the correlation function of C is given by

$$\begin{aligned}\langle C(t)C(t')\rangle &= \int d^3xW(\vec{x})\int d^3x'W(\vec{x}')\langle c(\vec{x},t)c(\vec{x}',t)\rangle &\text{(A297)}\\ &= \int d^3xW(\vec{x})\int d^3x'W(\vec{x}')\int\frac{d^3k}{(2\pi)^3}e^{i\vec{k}\cdot(\vec{x}-\vec{x}')}\int\frac{d\omega}{2\pi}e^{-i\omega(t-t')}S_c(\vec{k},\omega) &\text{(A298)}\\ &= \int\frac{d\omega}{2\pi}e^{-i\omega(t-t')}\int\frac{d^3k}{(2\pi)^3}|\tilde{W}(\vec{k})|^2 S_c(\vec{k},\omega), &\text{(A299)}\end{aligned}$$

where as usual $\tilde{W}$ denotes the Fourier transform of W. If we want to identify integration with a weight W as equivalent to computing an average over the volume V, then we must have $\tilde{W}(0) = 1$, and $\tilde{W}(\vec{k})$ must decay to zero for $k \gg 1/\ell$, where ℓ is the characteristic linear dimension of the region over which we are averaging.

Equation (A299) allows us to identify the power spectrum of fluctuations in C,

$$S_C(\omega) = \int\frac{d^3k}{(2\pi)^3}|\tilde{W}(\vec{k})|^2 S_c(\vec{k},\omega). \qquad \text{(A300)}$$

To make progress let's assume that the region we are averaging over is spherically symmetric, so that

$$\begin{aligned}S_C(\omega) &= \int\frac{d^3k}{(2\pi)^3}|\tilde{W}(\vec{k})|^2 S_c(\vec{k},\omega)\\ &= \bar{c}\int\frac{d^3k}{(2\pi)^3}|\tilde{W}(\vec{k})|^2\frac{2Dk^2}{\omega^2 + (Dk^2)^2} &\text{(A301)}\\ &= \bar{c}\frac{1}{(2\pi)^3}\int_0^\infty dk\,4\pi k^2|\tilde{W}(k)|^2\frac{2Dk^2}{\omega^2 + (Dk^2)^2}. &\text{(A302)}\end{aligned}$$

Now if we want the compute the variance in C, we have

$$\langle(\delta C)^2\rangle \equiv \int \frac{d\omega}{2\pi} S_C(\omega) \tag{A303}$$

$$= \bar{c}\frac{1}{(2\pi)^3}\int_0^\infty dk\, 4\pi k^2 |\tilde{W}(k)|^2 \int \frac{d\omega}{2\pi}\frac{2Dk^2}{\omega^2+(Dk^2)^2} \tag{A304}$$

$$= \bar{c}\frac{1}{2\pi^2}\int_0^\infty dk\, k^2|\tilde{W}(k)|^2. \tag{A305}$$

As an approximation, we can say that the effect of $|\tilde{W}(k)|^2$ is to cut the k integral off at $k \sim 2\pi/\ell$, in which case we have

$$\langle(\delta C)^2\rangle = \bar{c}\frac{1}{2\pi^2}\int_0^\infty dk\, k^2|\tilde{W}(k)|^2$$

$$\sim \bar{c}\frac{1}{2\pi^2}\int_0^{2\pi/\ell} dk\, k^2 \tag{A306}$$

$$\sim \frac{\bar{c}}{\ell^3}. \tag{A307}$$

Since $\bar{C} = \bar{c}$, we can also write this as

$$\frac{\langle(\delta C)^2\rangle}{\bar{C}^2} \sim \frac{1}{\bar{c}\ell^3}, \tag{A308}$$

and we recognize $N = \bar{c}\ell^3$ as the mean number of molecules in the sampling volume. Thus, the rigorous calculation from the fluctuation dissipation theorem gives us back our intuition about the fractional variance in concentration being $1/N$.

To get the rest of the Berg–Purcell result, let's go back to Eq (A302) and finish computing the power spectrum of C, in the same approximations:

$$S_C(\omega) = 2\bar{c}\frac{1}{(2\pi)^3}\int_0^\infty dk\, 4\pi k^2|\tilde{W}(k)|^2\frac{Dk^2}{\omega^2+(Dk^2)^2}$$

$$\sim \frac{\bar{c}}{\pi^2}\int_0^{2\pi/\ell} dk\,\frac{Dk^4}{\omega^2+(Dk^2)^2}. \tag{A309}$$

If ℓ is small, then the characteristic time for diffusion across the averaging volume, $\tau \sim \ell^2/D$, is also small, and hence any frequencies that are likely to be relevant for the cell's measurements of concentration are low compared with the scales on which $S_C(\omega)$ has structure. Thus we can confine our attention to the low frequency limit,

$$S_C(\omega \to 0) \sim \frac{\bar{c}}{\pi^2}\int_0^{2\pi/\ell} dk\,\frac{Dk^4}{(Dk^2)^2} = \frac{2\bar{c}}{\pi D\ell}. \tag{A310}$$

So we see that the concentration, averaged over a sampling volume of linear dimension ℓ has white noise in time. If we average over a time $\tau_{\rm avg}$, then we are sensitive to a bandwidth $1/\tau_{\rm avg}$, and we will see a variance

$$\langle(\delta C)^2\rangle_{\tau_{\rm avg}} \sim \frac{2\bar{c}}{\pi D\ell\tau_{\rm avg}}. \tag{A311}$$

Rewriting this as a fractional standard deviation, we have

$$\frac{\delta C_{\rm rms}}{\bar{C}} = \frac{1}{\bar{C}}\sqrt{\langle(\delta C)^2\rangle_{\tau_{\rm avg}}} \sim \left(\frac{2}{\pi}\right)^{1/2}\frac{1}{\sqrt{D\ell\bar{c}\tau_{\rm avg}}}, \tag{A312}$$

which is (except for the trivial factor $\sqrt{2/\pi}$) exactly the Berg–Purcell result.

Problem 187: Concentration fluctuations in one dimension. Repeat the analysis we have just done, but in one dimension. Before going through a detailed calculation, you should try to anticipate the answer. We still expect (from the $\sqrt{N}$ intuition) that $\langle(\delta C)^2\rangle \propto \bar{c}$, but since concentration has units of molecules per length in 1D, the other factors must be different. Try, for example,

$$\langle(\delta C)^2\rangle \sim \frac{\bar{c}}{(D\tau_{\rm avg})^n\ell^m}. \tag{A313}$$

How are n and m constrained by dimensional analysis? Can you argue, qualitatively, for particular values of these exponents? Finally, do the real calculation and get the analog of Eq (A311) in one dimension. Are you surprised by the role of ℓ (that is, by the value of m)? Can you explain why things come out this way?

Problem 188: Correlations seen by a moving observer. Generalize the discussion above to the case where the volume in which we measure the concentration is moving at speed v_0 in some direction. Provide a formula for the correlations across time in the observed noise. Show, in particular, that there is a correlation time $\tau_c \sim D/v_0^2$. How does this relate to the qualitative argument, discussed above, that bacteria must integrate for a minimum time $\sim D/v_0^2$ if they are to "outrun" diffusion?

So the Berg–Purcell argument certainly gives the right answer for the concentration fluctuations in a small volume. But biological systems don't actually count the molecules in a volume. Instead, the molecules bind to specific sites, and it is this binding which is detected, e.g. by activating an enzymatic reaction. The Berg–Purcell formula suggests that there is a limit to the accuracy of sensing or signaling that comes from the physics of diffusion alone, independent of these details. To see how this can happen, we need to analyze fluctuations in the binding of molecules to receptor sites, coupled to their diffusion. Let's start just with the binding events.

Consider a binding site for signaling molecules, and let the fractional occupancy of the site be n. If we do not worry about the discreteness of this one site, or about the fluctuations in concentration c of the signaling molecule, we can write a kinetic equation

$$\frac{dn(t)}{dt} = k_+c[1-n(t)] - k_-n(t). \tag{A314}$$

This describes the kinetics whereby the system comes to equilibrium, and the free energy F associated with binding is determined by detailed balance,

$$\frac{k_+c}{k_-} = \exp\left(\frac{F}{k_BT}\right). \tag{A315}$$

If we imagine that thermal fluctuations can lead to small changes in the rate constants, we can linearize Eq. (A314) to obtain

$$\frac{d\delta n}{dt} = -(k_+c + k_-)\delta n + c(1-\bar{n})\delta k_+ - \bar{n}\delta k_-. \quad \text{(A316)}$$

But from Eq. (A315) we have

$$\frac{\delta k_+}{k_+} - \frac{\delta k_-}{k_-} = \frac{\delta F}{k_B T}. \quad \text{(A317)}$$

Applying this constraint to Eq. (A316) we find that the individual rate constant fluctuations cancel and all that remains is the fluctuation in the thermodynamic binding energy δF:

$$\frac{d\delta n}{dt} = -(k_+c + k_-)\delta n + k_+c(1-\bar{n})\frac{\delta F}{k_B T}. \quad \text{(A318)}$$

Fourier transforming, we can solve Eq. (A318) to find the frequency dependent susceptibility of the coordinate n to its conjugate force F,

$$\tilde{\alpha}(\omega) \equiv \frac{\delta\tilde{n}(\omega)}{\delta\tilde{F}(\omega)} = \frac{1}{k_B T}\,\frac{k_+c(1-\bar{n})}{-i\omega + (k_+c + k_-)}. \quad \text{(A319)}$$

Now we can compute the power spectrum of fluctuations in the occupancy n using the fluctuation dissipation theorem,

$$\begin{aligned} S_n(\omega) &= \frac{2k_B T}{\omega}\mathrm{Im}\left[\frac{\delta\tilde{n}(\omega)}{\delta\tilde{F}(\omega)}\right] && \text{(A320)}\\ &= \frac{2k_+c(1-\bar{n})}{\omega^2 + (k_+c + k_-)^2}. && \text{(A321)}\end{aligned}$$

It is convenient to rewrite this as

$$S_n(\omega) = \langle(\delta n)^2\rangle\frac{2\tau_c}{1+(\omega\tau_c)^2}, \quad \text{(A322)}$$

where the total variance is

$$\begin{aligned} \langle(\delta n)^2\rangle &= \int\frac{d\omega}{2\pi}S_n(\omega) = k_B T\left.\frac{\delta\tilde{n}(\omega)}{\delta\tilde{F}(\omega)}\right|_{\omega=0} && \text{(A323)}\\ &= \frac{k_+c(1-\bar{n})}{k_+c + k_-} && \text{(A324)}\\ &= \bar{n}(1-\bar{n}), && \text{(A325)}\end{aligned}$$

and the correlation time is given by

$$\tau_c = \frac{1}{k_+c + k_-}. \quad \text{(A326)}$$

To make sense out of these results, remember what happens if we flip a coin that is biased to produce heads a fraction f of the time. On each trial we count either one or zero heads, so the mean count is f and the mean square count is also f; the variance is $f(1-f)$, exactly as in Eq (A325): when we check the occupancy of the receptor, the outcome is determined by the equivalent of flipping a biased coin, where the bias is determined by the Boltzmann distribution.

The Lorentzian form of the power spectrum in Eq (A322) is equivalent to an exponential decay of correlations,

$$\begin{aligned} \langle\delta n(t)\delta n(t')\rangle &= \int\frac{d\omega}{2\pi}e^{-i\omega(t-t')}S_n(\omega) && \text{(A327)}\\ &= \langle(\delta n)^2\rangle\int\frac{d\omega}{2\pi}e^{-i\omega(t-t')}\frac{2\tau_c}{1+(\omega\tau_c)^2} && \text{(A328)}\\ &= \langle(\delta n)^2\rangle e^{-|t-t'|/\tau_c}. && \text{(A329)}\end{aligned}$$

The exponential decay of correlations is what we expect when the transitions between the available states have no memory. To be precise about this, if we imagine that a system is in one state at time $t = 0$, and there is some constant probability per unit time k of transitions out of this state (with, in the simplest case, no returns to the initial state), then the probability $p(t)$ of still being in the initial state at time t must obey

$$\frac{dp(t)}{dt} = -kp(t), \quad \text{(A330)}$$

and hence $p(t) = e^{-kt}$. This intuition about the connection of exponential decays to the lack of memory is very general, and should remind you of the exponential distribution of times between transitions in the calculation of chemical reaction rates (Section II.A), and of the exponential distribution of times between events in a Poission process (see Appendix A.1). In the present context, the exponential decay of correlations tell us that the spontaneous transitions between the occupied and unoccupied states of the receptor occur with constant probability per unit time, or as Markovian jumps. The jumping rates are just the rates k_+ and k_-, which means that when we write chemical kinetic models for a whole ensemble of molecules, we also can interpret these as Markov models for transitions among the states of individual molecules in the ensemble.

It is interesting that we recover the results for Markovian jumping between two states without making this microscopic model explicit. All we assume is the macroscopic kinetics and that the system is in thermal equilibrium so that we can apply the fluctuation dissipation theorem. In principle many different microscopic models can describe the molecular phenomena that are at the basis of some observed macroscopic behavior, and we know that many aspects of behavior in thermal equilibrium are independent of these details. The statistics of fluctuations in a chemical kinetic system are an example of this, at least near equilibrium.

The good news, then, is that fluctuations in receptor occupancy are an inevitable consequence of the *macroscopic*, average behavior of receptor–ligand interactions, independent of hypothesis about molecular details. The bad news is that the form of the results doesn't seem very related to the ideas of Berg and Purcell about the precision of concentration measurements. To make these connections clear we need to couple the dynamics of receptor occupancy to the diffusion of the ligand.

When the concentration is allowed to fluctuate we write

$$\frac{dn(t)}{dt} = k_+ c(\vec{x}_0, t)[1 - n(t)] - k_- n(t), \tag{A331}$$

where the receptor is located at $\vec{x}_0$, and

$$\frac{\partial c(\vec{x}, t)}{\partial t} = D\nabla^2 c(\vec{x}, t) - \delta(\vec{x} - \vec{x}_0)\frac{dn(t)}{dt}. \tag{A332}$$

The first equation is as before, but with notation to remind us that the concentration c is dynamic. The second equation states that the ligand diffuses with diffusion constant D, and when the receptor located at $\vec{x}_0$ increases its occupancy it removes exactly one molecule from solution at that point.

Problem 189: Coupling diffusion and binding. In this problem you'll fill the details needed for the analysis of Eq's (A331) and (A332).

(a.) Begin by noticing that Eq (A332) is linear, so you should be able to solve it exactly. Use Fourier transforms, both in space and time, and then transform back to give a formal expression for

$$\tilde{c}(\vec{x}, \omega) = \int dt\, e^{+i\omega t} c(\vec{x}, t). \tag{A333}$$

(b.) Linearize Eq (A331), in the same way that we did in the preceding derivation, leading from Eq (A314) to (A319). Along the way you will need an expression for $\tilde{c}(\vec{x}_0, \omega)$, which you can take from (a.). When the dust settles, you should find Eq's (A334, A335)

Following the same steps as above, we find the linear response function

$$\frac{\delta\tilde{n}(\omega)}{\delta\tilde{F}(\omega)} = \frac{k_+ c(1 - \bar{n})}{k_B T}\,\frac{1}{-i\omega[1 + \Sigma(\omega)] + (k_+\bar{c} + k_-)} \tag{A334}$$

$$\Sigma(\omega) = k_+(1 - \bar{n})\int\frac{d^3k}{(2\pi)^3}\frac{1}{-i\omega + Dk^2} \tag{A335}$$

The "self–energy" $\Sigma(\omega)$ is ultraviolet divergent, which can be traced to the delta function in Eq (A332); we have assumed that the receptor is infinitely small. A more realistic treatment would give the receptor a finite size, which is equivalent to cutting off the k integrals at some (large) $\Lambda \sim \pi/a$, with a the linear dimension of the receptor. If we imagine mechanisms which read out the receptor occupancy average over a time τ long compared to the correlation time τ_c of the noise, then the relevant quantity is the low frequency limit of the noise spectrum. Hence,

$$\Sigma(\omega \ll D/a^2) \approx \Sigma(0) = \frac{k_+(1 - \bar{n})}{2\pi Da}, \tag{A336}$$

and

$$\frac{\delta\tilde{n}(\omega)}{\delta\tilde{F}(\omega)} = \frac{k_+\bar{c}(1 - \bar{n})}{k_B T}\left[-i\omega\left(1 + \frac{k_+(1 - \bar{n})}{2\pi Da}\right) + (k_+\bar{c} + k_-)\right]^{-1}, \tag{A337}$$

where $\bar{c}$ is the mean concentration. Applying the fluctuation–dissipation theorem once again we find the spectral density of occupancy fluctuations,

$$S_n(\omega) \approx 2k_+\bar{c}(1 - \bar{n})\frac{1 + \Sigma(0)}{\omega^2(1 + \Sigma(0))^2 + (k_+\bar{c} + k_-)^2}\,. \tag{A338}$$

The total variance in occupancy is unchanged, since this is an equilibrium property of the system, while coupling to concentration fluctuations serves only to change the kinetics.

Problem 190: Reading off the results. You should be able to verify the statements in the last sentence without detailed calculation. Explain how to "read" Eq (A338) and identify the total variance and correlation time.

Coupling to concentration fluctuations does serve to renormalize the correlation time of the noise,

$$\tau_c \to \tau_c[1 + \Sigma(0)]. \tag{A339}$$

The new τ_c can be written as

$$\tau_c = \frac{1 - \bar{n}}{k_-} + \frac{\bar{n}(1 - \bar{n})}{2\pi Da\bar{c}}, \tag{A340}$$

so there is a lower bound on τ_c, independent of the kinetic parameters $k_\pm$,

$$\tau_c > \frac{\bar{n}(1-\bar{n})}{2\pi D a \bar{c}}. \tag{A341}$$

As discussed previously, the relevant quantity is the low frequency limit of the noise spectrum,

$$S_n(\omega = 0) = 2k_+\bar{c}(1-\bar{n}) \cdot \frac{1+\Sigma(0)}{(k_+\bar{c}+k_-)^2} \tag{A342}$$

$$= \frac{2\bar{n}(1-\bar{n})}{k_+\bar{c}+k_-} + \frac{[\bar{n}(1-\bar{n})]^2}{\pi D a \bar{c}}. \tag{A343}$$

If we average for a time τ, then the root-mean-square error in our estimate of n will be

$$\delta n_{\rm rms} = \sqrt{S_n(0) \cdot \frac{1}{\tau}}, \tag{A344}$$

and we see that this noise level has a minimum value independent of the kinetic parameters $k_\pm$,

$$\delta n_{\rm rms} > \frac{\bar{n}(1-\bar{n})}{\sqrt{\pi D a \bar{c} \tau}}. \tag{A345}$$

To relate these results back to the discussion by Berg and Purcell, we note that the $\omega = 0$ response of the mean occupancy to changes in concentration can be written as

$$\frac{d\bar{n}}{d\ln c} = \bar{n}(1-\bar{n}). \tag{A346}$$

Thus, the fluctuations in n are equivalent to fluctuations in c:

$$\frac{\delta c_{\rm eff}}{\bar{c}} = (\delta \ln c)_{\rm eff} = \delta n_{\rm rms} \left(\frac{d\bar{n}}{d\ln c}\right)^{-1} = \frac{1}{\sqrt{\pi D a \bar{c} \tau}}. \tag{A347}$$

Except for the factor of $\sqrt{\pi}$, this is the Berg–Purcell result once again.

A startling feature of the Berg–Purcell argument is that (it seems) it can be used both when a is the size of a single receptor molecule and when a is the size of the entire bacterium. Naively, we might expect that if there are N receptors on the surface of the cell, then the signal–to–noise ratio for concentration measurements should be N times better, and correspondingly the threshold for reliable detection should be $\sqrt{N}$ times smaller,

$$\frac{\delta c_{\rm eff}}{\bar{c}} \sim \frac{1}{\sqrt{DNa\bar{c}\tau}}. \tag{A348}$$

On the other hand, if we use the Berg–Purcell limit and take the linear dimensions of the detector to be the radius R of the bacterium, we should obtain

$$\frac{\delta c_{\rm eff}}{\bar{c}} \sim \frac{1}{\sqrt{DR\bar{c}\tau}}. \tag{A349}$$

What is going on? Does something special happen when $N \sim R/a$, so there is a crossover between the two results?

If we imagine a very large cell, and place $N = 2$ two receptors on opposite sides of the cell surface, it is hard to imagine that there is anything wrong with the argument leading to Eq (A348). More generally, if the receptors are far apart, it is very plausible that they report independent measurements of the concentration, and so Eq (A348) should be correct. On the other hand, if we imagine bringing two receptors closer and closer together, at some point they will start to interact—a molecule released from one receptor can diffuse over and bind to the other receptor—and this interaction might lead to correlations in the noise, and a break down of the simple $\sqrt{N}$ improvement in the threshold for reliable detection.

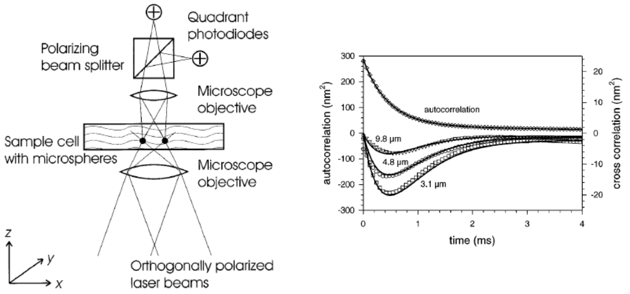

FIG. 174 Correlated Brownian motion, from Meiners & Quake (1999). At left, a schematic of the experiment. The laser beams from the bottom of the figure create two optical traps, which hold the microspheres in approximately harmonic potential wells. The optics at the top allow for measurements of the spheres' positions with nanometer precision. At right, measurements of the auto– and cross–correlations of the spheres' positions; different curves for the cross–correlation correspond to different mean separations of the particles, which is expected to modulate the coupling between them through the fluid.

To understand how diffusive interactions lead to correlations among receptors, it is useful to think about a simpler problem. Suppose that we have two balls in a fluid. If they are very far apart, each one experiences a drag force and undergoes Brownian motion, and the Brownian fluctuations in the position of each ball are independent of those in the other. If we bring the two balls together, however, we know that they can influence each other through the fluid: If one ball moves at velocity v_1 it not only experiences a drag force $-\gamma v_1$, it also applies a "coupling" force $\gamma_c(v_1 - v_2)$ to the other ball (which may be moving at velocity v_2; clearly if $v_1 = v_2$ there should be no coupling force). If the balls are close enough that γ_c is significant, then in fact the Brownian motions of the two balls become correlated. This correlation can be derived from the fluctuation–dissipation theorem, and it also makes intuitive sense since a random Brownian step of one object applies a force to the other. We can also see this effect experimentally, as in Fig 174.

Problem 191: Correlated Brownian motion. To make the situation in the previous paragraph precise, consider the case where the particles are bound by springs (so they can't diffuse away from each other and reduce the coupling). Then, in the overdamped case, the equations of motion are

$$\gamma\frac{dx_1}{dt} = -\kappa x_1 - \gamma_c\left(\frac{dx_1}{dt} - \frac{dx_2}{dt}\right) + F_1(t) \tag{A350}$$

$$\gamma\frac{dx_2}{dt} = -\kappa x_2 - \gamma_c\left(\frac{dx_2}{dt} - \frac{dx_1}{dt}\right) + F_1(t) \tag{A351}$$

where κ is the stiffness of the springs (assumed identical, for simplicity), and $F_{\rm i}(t)$ is an external force applied to each particle i.

(a.) Derive the linear response function matrix, $\tilde{\alpha}_{\rm ij}(\omega)$ such that

$$\tilde{x}_{\rm i}(\omega) = \sum_{\rm j}\tilde{\alpha}_{\rm ij}(\omega)\tilde{F}_{\rm j}(\omega). \tag{A352}$$

(b.) The generalization of the fluctuation dissipation theorem to many degrees of freedom states that the "cross–spectrum" of variables i and j, defined by

$$\langle x_{\rm i}(t)x_{\rm j}(t')\rangle = \int\frac{d\omega}{2\pi}e^{-i\omega(t-t')}S_{\rm ij}(\omega), \tag{A353}$$

is given by

$$S_{\rm ij}(\omega) = \frac{2k_BT}{\omega}{\rm Im}\left[\tilde{\alpha}_{\rm ij}(\omega)\right]. \tag{A354}$$

Use this to derive the cross–spectrum of the position fluctuations for the two particles.

(c.) Despite the viscous coupling, the potential energy is just the sum of contributions from the two particles. From the Boltzmann distribution, then, the positions should be independent variables. Use your results in (b) to show that $\langle x_{\rm i}x_{\rm j}\rangle = \delta_{\rm ij}k_BT/\kappa$. Notice that this corresponds to the *instantaneous* positions of the particles, as we would measure by taking a snapshot (with a fast camera).

(d.) Suppose that instead of taking snapshots of the positions, we average (as in the discussion above) for a long time, so what is relevant is the low frequency limit of the power spectra. Show that now the correlations are nonzero, and give an explicit formula for the covariance matrix of fluctuations in the temporally averaged positions.

If we imagine that positions of the Brownian particles are like receptor occupancies, and an applied force on all the particles is like a change in concentration of the relevant ligand, then diffusion of the ligand serves the same coupling effect as the viscosity of the fluid and will generate correlations among the occupancy fluctuations of nearby receptors. These correlations mean that using the positions or velocities of N Brownian particles to infer the applied force is *not* $\sqrt{N}$ more accurate than using one particle, and similarly using N receptors will not generate a concentration measurement that is $\sqrt{N}$ times more accurate than is obtained with one receptor.

If we have N receptors, each of size a arrayed on a structure of linear dimension R such as a ring or a sphere, then as N becomes large the receptors are coming closer and closer together, and we expect that correlations become stronger. If we have two detectors making measurements with noise that becomes more and more strongly correlated, at some point they start to act like one big detector. If we work through the details of the calculations for the case of multiple receptors,[98] indeed we find that as N become large, the correlations among the different receptors become limiting, and the threshold for reliable detection approaches Eq (A349): the $N \to \infty$ receptors packed into a structure with linear dimension R acts like one receptor of size R. If we go back to the intuitive Berg–Purcell argument about counting molecules in a volume and getting a fresh count each time the volume clears from diffusion, what this means is that packing many receptor sites into a region of size R eventually means that we get to count the molecules in a volume $\sim R^3$. There are geometrical factors for different spatial arrangements of the receptors, but like the $\sqrt{\pi}$ in Eq (A347) these aren't a big deal.

Almost all of PhD students in physics have seen some cases of the fluctuation dissipation theorem, somewhere in their statistical mechanics courses. Whether you have seen the general formulation depends a bit on who taught the course, and how far you went. As usual, an excellent discussion can be found in Landau & Lifshitz (1977). A later volume in the Landau and Lifshitz series (Lifshitz & Pitaevskii 1980) provides a clear discussion of concentration fluctuations, in Section 89. Many people find the idea of correlations between Brownian particles to be surprising, so it's worth looking at real experiments that measure these correlations (Meiners & Quake 1999). [should add refs to measurements on concentration fluctuations—Feher and Weissman?]

Landau & Lifshitz 1977: *Statistical Physics.* LD Landau & EM Lifshitz (Pergamon, Oxford, 1977).

Lifshitz & Pitaevskii 1980: *Statistical Physics, Part 2.* EM Lifshitz & LP Pitaevskii (Pergamon, Oxford, 1980).

Meiners & Quake 1999: Direct measurement of hydrodynamic cross correlations between two particles in an external potential. JC Meiners & SR Quake, *Phys Rev Lett* **82,** 2211–2214 (1999).

The idea that fluctuations in certain chemical systems could be described using the fluctuation dissipation theorem must have occurred to many people, and I remember discussing it long ago (Bialek 1987). The emergence of experiments on noise in the control of gene expression made it more interesting to get everything straight, so my colleagues and I did this in a series of papers (Bialek & Setayeshgar 2005, 2008; Tkačik & Bialek 2009).

Bialek 1987: Physical limits to sensation and perception. W Bialek, *Ann Rev Biophys Biophys Chem* **16,** 455–478 (1987).

[98] See the references at the end of this section for details.

Bialek & Setayeshgar 2005: Physical limits to biochemical signaling. W Bialek & S Setayeshgar, *Proc Nat'l Acad Sci (USA)* **102,** 10040–10045 (2005).

Bialek & Setayeshgar 2008: Cooperativity, sensitivity and noise in biochemical signaling. W Bialek & S Setayeshgar, *Phys Rev Lett* **100,** 258101 (2008).

Tkačik & Bialek 2009: Diffusion, dimensionality and noise in transcriptional regulation. G Tkačik & W Bialek, *Phys Rev E* **79,** 051901 (2009).

7. Dimensionality reduction

I am leaving this unwritten for now, in the interests of getting something readable out more quickly. I think it is straightforward to write. To give a sense of what goes here, I have started to compile the references. Evidently papers by my colleagues and myself are over–represented; of course a full account will look at a broader literature.

Problem 192: Analysis of a sensory neuron. [Get a big data set from Rob on H1, and use it to take the students through reverse correlation, spike triggered covariance and (maybe) maximally informative dimensions. Have repeats so one can compare reduced models with the real information per spike.]

Problem 193: Analysis of DNA sequences. [Get data from Justin and take the students through a small version of the problem (maybe the RNAP site).]

d'Avella & Bizzi 1998: Low dimensionality of surpaspinally induced force fields. A d'Avella & E Bizzi, *Proc Nat'l Acad Sci (USA)* **95,** 7711–7714 (1998).

Bialek & de Ruyter van Steveninck 2005: Features and dimensions: Motion estimation in fly vision. W Bialek & R de Ruyter van Steveninck, arXiv:q–bio/0505003 (2005).

de Boer & Kuyper 1968: Triggered correlation. E de Boer & P Kuyper, *IEEE Trans Biomed Eng* **15,** 169–179 (1968).

Chigirev & Bialek 2004: Optimal manifold representation of data: An information theoretic perspective. DV Chigirev & W Bialek, in *Advances in Neural Information Processing 16,* S Thrun, L Saul & B Schölkopf, eds, pp 161–168 (MIT Press, Cambridge, 2004).

Fairhall et al 2006: Selectivity for multiple stimulus features in retinal ganglion cells. AL Fairhall, CA Burlingame, R Narasimhan, RA Harris, JL Puchalla & MJ Berry II, *J Neurophysiol* **96,** 2724–2738 (2006).

Kinney et al 2007: Precise physical models of protein–DNA interaction from high-throughput data. JB Kinney, G Tkačik & CG Callan Jr, *Proc Natl Acad Sci (USA)* **104,** 501–506 (2007).

Kinney et al 2010: Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. JB Kinney, A Murugan, CG Callan Jr & EC Cox, *Proc Nat'l Acad Sci (USA)* **107,** 9158–9163 (2010).

Osborne et al 2005: A sensory source for motor variation. LC Osborne, SG Lisberger & W Bialek, *Nature* **437,** 412–416 (2005).

Rust et al 2005: Spatiotemporal elements of macaque V1 receptive fields. NC Rust , O Schwartz, JA Movshon & EP Simoncelli, *Neuron* **46,** 945–956 (2005).

de Ruyter van Steveninck & Bialek 1988: Real–time performance of a movement sensitive neuron in the blowfly visual system: Coding and information transfer in short spike sequences. R de Ruyter van Steveninck & W Bialek, *Proc R. Soc London Ser. B* **234,** 379–414 (1988).

Sanger 2000: Human arm movements described by a low–dimensional superposition of principal components. TD Sanger, *J Neurosci* **20,** 1066–1072 (2000).

Sharpee et al 2004: Analyzing neural responses to natural signals: Maximally informative dimensions. T Sharpee, NC Rust & W Bialek, *Neural Comp* **16,** 223–250 (2004); arXiv:physics/0212110 (2002).

Stephens et al 2008: Dimensionality and dynamics in the behavior of *C. elegans*. GJ Stephens, B Johnson–Kerner, W Bialek & WS Ryu, *PLoS Comp Bio* **4,** e1000028 (2008); arXiv:0705.1548 [q–bio.OT] (2007).

Stephens et al 2011: Searching for simplicity in the analysis of neurons and behavior. GJ Stephens, LC Osborne & W Bialek, *Proc Nat'l Acad Sci (USA)* in press (2011); arXiv.org:1012.3896 [q–bio.NC] (2010).

8. Maximum entropy

This section is a bit old. It needs to be revised in light of what happens in Sections III.A and III.D. Be sure that we go into RESULTS on these methods, as promised for neurons, at least.

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

$$\langle f_{\mathrm{i}} \rangle = \sum_D P(D) f_{\mathrm{i}}(D), \tag{A355}$$

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

$$\langle f_0 \rangle = \sum_D P(D) = 1. \tag{A356}$$

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

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

$$\tilde{S}[P(D)] = -\sum_D P(D)\log_2 P(D) - \sum_{\mathrm{i}=0}^{K} \tilde{\lambda}_\mathrm{i} \langle f_\mathrm{i} \rangle \tag{A357}$$

$$= -\frac{1}{\ln 2}\sum_D P(D)\ln P(D) - \sum_{\mathrm{i}=0}^{K} \lambda_\mathrm{i} \sum_D P(D) f_\mathrm{i}(D). \tag{A358}$$

Our problem, then, is to find the maximum of the function $\tilde{S}$, but this is easy because the probability for each value of D appears independently. As usual, we differentiate and set the result to zero:

$$0 = \frac{\partial \tilde{S}}{\partial P(D)} = -\frac{1}{\ln 2}\left[\ln P(D) + 1\right] - \sum_{\mathrm{i}=0}^{K} \tilde{\lambda}_\mathrm{i} f_\mathrm{i}(D). \tag{A359}$$

Rearranging, we have

$$\ln P(D) = -1 - \sum_{\mathrm{i}=0}^{K} (\ln 2)\tilde{\lambda}_\mathrm{i} f_\mathrm{i}(D) \tag{A360}$$

$$P(D) = \frac{1}{Z}\exp\left[-\sum_{\mathrm{i}=1}^{K} \lambda_\mathrm{i} f_\mathrm{i}(D)\right], \tag{A361}$$

where $\lambda_\mathrm{i} = (\ln 2)\tilde{\lambda}_\mathrm{i}$, and $Z = \exp(1 + \lambda_0)$ is a normalization constant. Notice that this gives us the *form* of the maximum entropy distribution, but we still have to adjust the constants $\{\lambda_\mathrm{i}\}$ so that the distribution $P(D)$ predicts the measured values of the expectation values in Eq (A355).

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

$$\langle E \rangle = \sum_\mathrm{n} P_\mathrm{n} E_\mathrm{n}, \tag{A362}$$

then the maximum entropy distribution is

$$P_\mathrm{n} = \frac{1}{Z}\exp(-\lambda E_\mathrm{n}), \tag{A363}$$

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

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

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

$$\bar{f}_1 = \langle x \rangle = \int dx P(x) x, \quad \text{and} \tag{A364}$$

$$\bar{f}_2 = \langle x^2 \rangle = \int dx P(x) x^2, \tag{A365}$$

so we have $f_1(x) = x$ and $f_2(x) = x^2$. Thus, from Eq. (A361), the maximum entropy distribution is of the form

$$P(x) = \frac{1}{Z}\exp(-\lambda_1 x - \lambda_2 x^2). \tag{A366}$$

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

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x - \langle x\rangle)^2\right]. \qquad \text{(A367)}$$

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

Problem 194: Less than maximum entropy. Many natural signals are strongly nonGaussian. In particular exponential (or nearly exponential) distribution are common in studies on the statistics of natural images and natural sounds. With the same mean (which you can call zero) and variance, what is the difference in entropy between the exponential [$P(x) \propto \exp(-\lambda|x|)$] and Gaussian distributions? If we imagine that this difference is relevant to every pixel (or to every Fourier component) in an image, is this significant compared to the 8 bits/pixel of a standard digital image? What if $P(x) \propto \exp(-\lambda|x|^\mu)$, with $\mu < 1$?

[maybe we should put the start of networks here?]

[maybe this should be connections, more generally (including what we have to say about counting), and that would leave a section to address the experimental situation more specifically?]

Probability distributions that have the maximum entropy form of Eq. (A361) are special not only because of their connection to statistical mechanics, but because they form what the statisticians call an 'exponential family,' which seems like an obvious name. The important point is that exponential families of distributions are (almost) unique in having sufficient statistics. To understand what this means, consider the following problem: we observe a set of samples $D_1, D_2, \cdots, D_N$, each of which is drawn independently and at random from a distribution $P(D|\{\lambda_{\rm i}\})$. Assume that we know the form of this distribution but not the values of the parameters $\{\lambda_{\rm i}\}$. How can we estimate these parameters from the set of observations $\{D_{\rm n}\}$? Notice that our data set $\{D_{\rm n}\}$ consists of N numbers, and N can be very large; on the other hand there typically are a small number $K \ll N$ of parameters $\lambda_{\rm i}$ that we want to estimate. Even in this limit, no finite amount of data will tell us the exact values of the parameters, and so we need a probabilistic formulation: we want to compute the distribution of parameters given the data, $P(\{\lambda_{\rm i}\}|\{D_{\rm n}\})$. We do this using Bayes' rule,

$$P(\{\lambda_{\rm i}\}|\{D_{\rm n}\}) = \frac{1}{P(\{D_{\rm n}\})} \cdot P(\{D_{\rm n}\}|\{\lambda_{\rm i}\})P(\{\lambda_{\rm i}\}), \qquad \text{(A368)}$$

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

$$P(\{D_{\rm n}\}|\{\lambda_{\rm i}\}) = \prod_{\rm n=1}^{N} P(D_{\rm n}|\{\lambda_{\rm i}\}). \qquad \text{(A369)}$$

For probability distributions of the maximum entropy form we can proceed further, using Eq. (A361):

$$\begin{aligned} P(\{\lambda_{\rm i}\}|\{D_{\rm n}\}) &= \frac{1}{P(\{D_{\rm n}\})} \cdot P(\{D_{\rm n}\}|\{\lambda_{\rm i}\})P(\{\lambda_{\rm i}\}) \\ &= \frac{P(\{\lambda_{\rm i}\})}{P(\{D_{\rm n}\})} \prod_{\rm n=1}^{N} P(D_{\rm n}|\{\lambda_{\rm i}\}) && \text{(A370)} \\ &= \frac{P(\{\lambda_{\rm i}\})}{Z^N P(\{D_{\rm n}\})} \prod_{\rm n=1}^{N} \exp\left[-\sum_{\rm i=1}^{K} \lambda_{\rm i} f_{\rm i}(D_{\rm n})\right] && \text{(A371)} \\ &= \frac{P(\{\lambda_{\rm i}\})}{Z^N P(\{D_{\rm n}\})} \exp\left[-N \sum_{\rm i=1}^{K} \lambda_{\rm i} \frac{1}{N} \sum_{\rm n=1}^{N} f_{\rm i}(D_{\rm n})\right]. && \text{(A372)} \end{aligned}$$

We see that *all* of the information that the data points $\{D_{\rm n}\}$ can give about the parameters $\lambda_{\rm i}$ is contained in the average values of the functions f_i over the data set, or the 'empirical means' $\bar{f}_{\rm i}$,

$$\bar{f}_{\rm i} = \frac{1}{N}\sum_{\rm n=1}^{N} f_{\rm i}(D_{\rm n}). \quad \text{(A373)}$$

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

$$P(\{\lambda_{\rm i}\}|D_1, D_2, \cdots, D_N) = P(\{\lambda_{\rm i}\}|\{\bar{f}_{\rm j}\}), \quad \text{(A374)}$$

and a consequence of this is the information theoretic statement

$$I(D_1, D_2, \cdots, D_N \to \{\lambda_{\rm i}\}) = I(\{\bar{f}_{\rm j}\} \to\{\ \lambda_{\rm i}\}). \quad \text{(A375)}$$

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

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

[say more about this!]

[where do we put connection of matching expectation values to maximum likelihood?]

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

Consider the (Boltzmann) probability distribution for the states of a system in thermal equilibrium. If we expand the Hamiltonian as a sum of terms (operators) then the family of possible probability distributions is an exponential family in which the coupling constants for each operator are the parameters analogous to the $\lambda_{\rm i}$ above. In principle there could be an infinite number of these operators, but for a given class of systems we usually find that only a finite set are "relevant" in the renormalization group sense: if we write an effective Hamiltonian for coarse grained degrees of freedom, then only a finite number of terms will survive the coarse graining procedure. If we have only a finite number of terms in the Hamiltonian, then the family of Boltzmann distributions has sufficient statistics, which are just the expectation values of the relevant operators. This means that the expectation values of the relevant operators carry all the information that the (coarse grained) configuration of the system can provide about the coupling constants, which in turn is information about the identity or microscopic structure of the system. Thus the statement that there are only a finite number of relevant operators is also the statement that a finite number of expectation values carries all the information about the microscopic dynamics. The 'if' part of this statement is obvious: if there are only a finite number of relevant operators, then the expectation values of these operators carry all the information about the identity of the system. The statisticians, through the theorem about the uniqueness of exponential families, give us the 'only if': a finite number of expectation values (or correlation functions) can provide all the information about the system *only if* the effective Hamiltonian has a finite number of relevant operators. I suspect that there is more to say along these lines.

An important example of the maximum entropy idea arises when the data D are generated by counting. Then the relevant variable is an integer $n = 0, 1, 2, \cdots$, and it is natural to imagine that what we know is the mean count $\langle n\rangle$. One way this problem can arise is that we are trying to communicate and are restricted to sending discrete or quantized units. An obvious case is in optical communication, where the quanta are photons. In the brain, quantization abounds: most neurons do not generate continuous analog voltages but rather communicate with one another through stereotyped pulses or spikes, and even if the voltages vary continuously transmission across a synapse involves the release of a chemical transmitter which is packaged into discrete vesicles. It can be relatively easy to measure the mean rate at which discrete events are counted, and we might want to know what bounds this mean rate places on the ability of the cells to convey information. Alternatively, there is an energetic cost associated with these discrete events—generating the electrical currents that underlie the spike, constructing and filling the vesicles, ... —and we might want to characterize the mechanisms by their cost per bit rather than their cost per event [Laughlin et al 1998, Sarpeshkar 1998].

If we know the mean count, there is (as for the Boltzmann distribution) only one function $f_1(n) = n$ that can appear in the exponential of the distribution, so that

$$P(n) = \frac{1}{Z}\exp(-\lambda n). \tag{A376}$$

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

$$S_{\rm max}(\text{counting}) = \log_2(1 + \langle n\rangle) + \langle n\rangle \log_2(1 + 1/\langle n\rangle). \tag{A377}$$

The information conveyed by counting something can never exceed the entropy of the distribution of counts, and if we know the mean count then the entropy can never exceed the bound in Eq. (A377). Thus, if we have a system in which information is conveyed by counting discrete events, the simple fact that we count only a limited number of events (on average) sets a bound on how much information can be transmitted. We will see that real neurons and synapses approach this fundamental limit.

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

$$S_{\rm max}(\text{counting}; \langle n\rangle \to \infty) \sim \log_2(\langle n\rangle e), \tag{A378}$$

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

$$\frac{1}{\langle n\rangle} S_{\rm max}(\text{counting}; \langle n\rangle << 1) \sim \log_2\left(\frac{e}{\langle n\rangle}\right), \tag{A379}$$

which is arbitrarily large for small mean count. This makes sense: rare events have an arbitrarily large capacity to surprise us and hence to convey information. It is important to note, though, that the maximum entropy per event is a monotonically decreasing function of the mean count. Thus if we are counting spikes from a neuron, counting in larger windows (hence larger mean counts) is always less efficient in terms of bits per spike.

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

$$S = -\sum_{N=0}^{\infty}\int d^N t_n P(t_1, t_2, \cdots, t_N)\log_2 P(t_1, t_2, \cdots, t_N) \tag{A380}$$

$$= \sum_{N=0}^{\infty} P(N) S_{\rm time}(N) - \sum_{N=0}^{\infty} P(N)\log_2 P(N), \tag{A381}$$

where we have made use of

$$P(t_1, t_2, \cdots, t_N) = P(t_1, t_2, \cdots, t_N|N)P(N), \tag{A382}$$

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

$$S_{\rm time}(N) = -\int d^N t_n P(t_1, t_2, \cdots, t_N|N)\log_2 P(t_1, t_2, \cdots, t_N|N). \tag{A383}$$

If all we fix is the mean count, $\langle N\rangle = \sum_N P(N)N$, then the conditional distributions for the locations of the events given the total number of events, $P(t_1, t_2, \cdots, t_N|N)$, are unconstrained. We can maximize the contribution of each of these terms to the entropy [the terms in the first sum of Eq. (A381)] by making the distributions $P(t_1, t_2, \cdots, t_N|N)$ uniform, but it is important to be careful about normalization. When we integrate over all the times $t_1, t_2, \cdots, t_N$, we are forgetting that the events are all identical, and hence that permutations of the times describe the same events. Thus the normalization condition is *not*

$$\int_0^T dt_1 \int_0^T dt_2 \cdots \int_0^T dt_N P(t_1, t_2, \cdots, t_N|N) = 1, \tag{A384}$$

but rather

$$\frac{1}{N!}\int_0^T dt_1 \int_0^T dt_2 \cdots \int_0^T dt_N P(t_1, t_2, \cdots, t_N|N) = 1. \tag{A385}$$

This means that the uniform distribution must be

$$P(t_1, t_2, \cdots, t_N|N) = \frac{N!}{T^N}, \tag{A386}$$

and hence that the entropy [substituting into Eq. (A381)] becomes

$$S = -\sum_{N=0}^{\infty} P(N)\left[\log_2\left(\frac{N!}{T^N}\right) + \log_2 P(N)\right]. \tag{A387}$$

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

$$\tilde{S} = -\sum_{N=0}^{\infty} P(N)\left[\log_2\left(\frac{N!}{T^N}\right) + \log_2 P(N) + \lambda_0 + \lambda_1 N\right], \tag{A388}$$

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

$$0 = \frac{\partial \tilde{S}}{\partial P(N)} \tag{A389}$$

$$= -\frac{1}{\ln 2}\left[\ln\left(\frac{N!}{T^N}\right) + \ln P(N) + 1\right] - \lambda_0 - \lambda_1 N, \tag{A390}$$

$$\ln P(N) = -\ln\left(\frac{N!}{T^N}\right) - (\lambda_1 \ln 2)N - (1 + \lambda_0 \ln 2) \tag{A391}$$

Combining terms and simplifying, we have

$$P(N) = \frac{1}{Z}\frac{(\lambda T)^N}{N!}, \tag{A392}$$

$$Z = \sum_{N=0}^{\infty} \frac{(\lambda T)^N}{N!} = \exp(\lambda T). \tag{A393}$$

This is the Poisson distribution.

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

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

[obviously this needs to be much better!] One important example is provided by binary strings. If we label 1s by spin up and 0s by spin down, the binary string is equivalent to an Ising chain $\{\sigma_\mathrm{i}\}$. Fixing the probability of a 1 is the same as fixing the mean magnetization $\langle\sigma_\mathrm{i}\rangle$. If, in addition, we specify the joint probability of two 1s occurring in bins separated by n steps (for all n), this is equivalent to fixing the spin–spin correlation function $\langle\sigma_\mathrm{i}\sigma_\mathrm{j}\rangle$. For simplicity, consider the case where the system is translation invariant, so the average magnetization is the same at all sites and the correlation function $\langle\sigma_\mathrm{i}\sigma_\mathrm{j}\rangle$ depends only on $\mathrm{i}-\mathrm{j}$. The maximum entropy distribution consistent with these constraints is an Ising model,

$$P[\{\sigma_\mathrm{i}\}] = \frac{1}{Z}\exp\left[-h\sum_\mathrm{i}\sigma_\mathrm{i} - \sum_\mathrm{ij} J(\mathrm{i}-\mathrm{j})\sigma_\mathrm{i}\sigma_\mathrm{j}\right]; \tag{A394}$$

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

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

[Where did this come from?] Before closing the discussion of maximum entropy distributions, note that our simple solution to the problem, Eq. (A361), might not work. Taking derivatives and setting them to zero works only if the solution to our problem is in the interior of the domain allowed by the constraints. It is also possible that the solution lies at the boundary of this allowed region. This seems especially likely when we combine different kinds of constraints, such as trying to find the maximum entropy distribution of images consistent both with the two–point correlation function and with the histogram of intensity values at one point. The relevant distribution is a 2D field theory with a (generally nonlocal) quadratic 'kinetic energy' and some arbitrary local potential; it is not clear that all combinations of correlations and histograms can be realized, nor that the resulting field theory will be stable under renormalization; the empirical histograms of local quantities in natural images *are* stable under renormalization [Ruderman and Bialek 1994]. There are many open questions here.

[why?]

9. Measuring information transmission

When we study classical mechanics, we can make a direct connection between the positions and momenta that appear in the equations of motion and the positions and momenta of the particles that we "see," as in the planetary orbits. This connection is a little bit subtle, since we don't actually measure particle positions; more likely we count the photons arriving at some detector, forming an image, or we measure the delay in propagation of a pulse used in radar, or ... . But one can think of classical mechanics, in contrast to quantum mechanics, as being the domain of physics in which these subtleties are not important. When we move to statistical physics the connection between what we write in equations and what we observe in the world becomes more abstract. The fundamental objects in statistical physics are probability distributions, and as a matter of definition one *cannot* measure a distribution. Instead, Nature (or even a controlled experiment) provides us with samples taken out of these distributions. This has very serious consequences for any attempt to "measure" information flow.

In thermodynamics, entropy changes are connected to heat flow, and so we can at least measure the difference in entropy between two states by tracking these heat flows. Indeed, there is a long tradition of integrating these changes from some convenient reference to "measure" the entropy of states at intermediate temperatures. As far as I know, there is no analog of this in the information theoretic context. Thus, although Shannon tells us that the entropy is a fundamental property of the distribution out which signals are drawn, there is no universal entropy meter.

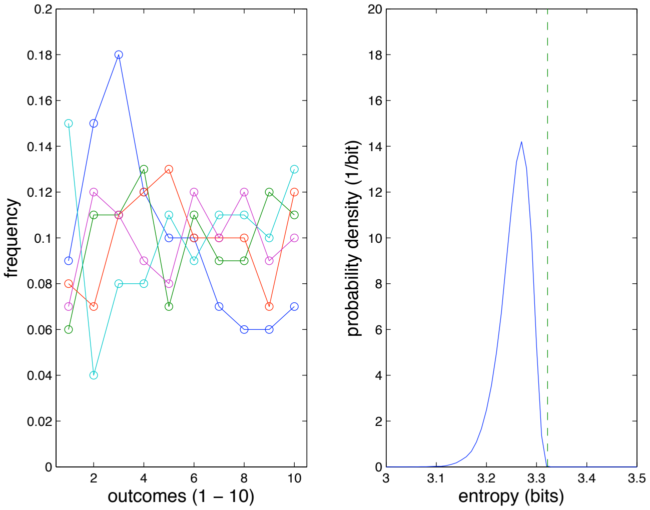

FIG. 175 The sampling problem in entropy estimation. At left, the frequency of occurrence found from five examples of $N = 100$ samples drawn out of $K = 10$ bins; the true probability distribution is flat, $p_{\rm i} = 0.1$ for all i. At right, we estimate the entropy of the distribution by identifying the observed frequencies with probabilities. The distribution of entropies obtained in this way, from many "experiments" with $N = 100$ and $K = 10$, is shown as a solid line, and should be compared with the true entropy, shown by a dashed line at $S_{\rm true} = \log_2(10)$.

To get a feeling for the problem, consider Fig 175. Here we have a variable that can take on ten possible values $({\rm i} = 1, 2, \cdots, 10)$, all equally likely ($p_{\rm i} = 0.1$ for all i), and we draw $N = 100$ samples. If we look at the frequency with which each possibility occurs, of course we don't see an exactly flat distribution. Since with 10 bins and 100 samples we expect 10 samples per bin, it's not surprising that the fluctuations are on the scale of $1/\sqrt{10} \sim 30\%$. These fluctuations, however, are random—they average to zero if we do the same experiment many times. The problem is that if we identify the frequencies of occurrence as our best estimates of the underlying probabilities, and use these estimates to compute the entropy, we make a systematic error, as is clear from the results in the right panel of Fig 175.

Problem 195: Experiment with sampling. Generate the analog of Fig 175 but with different values for the number of possible values K, where $\mathrm{i} = 1, 2, \cdots, K$. You should also try different probability distributions (e.g., $p_\mathrm{i} \propto 1/\mathrm{i}$, Zipf's law). Experiment. Convince yourself that, by identifying probabilities with the observed frequencies of occurrence, you always underestimate the entropy.

The problem illustrated in Fig 175 might seem very specific to the conditions of that simulation (e.g., that the true distribution is flat, and hence the entropy is maximal, so perhaps all errors have to be biased downward?), but in fact is very general. Let's consider drawing samples out of a discrete set of possibilities, $\mathrm{i} = 1, 2, \cdots, K$, with probabilities $\mathbf{p} \equiv \{p_1, p_2, \cdots, p_K\}$. If we draw N samples all together, we will find n_i examples of the outcome i, and of course on average $\langle n_\mathrm{i}\rangle = Np_\mathrm{i}$. Since we're counting random events, we expect that the variance of the number of events of type i will be equal to the mean, $\langle(\delta n_\mathrm{i})^2\rangle = \langle n_\mathrm{i}\rangle = Np_\mathrm{i}$. If we define the frequency of events in the usual way as $f_\mathrm{i} = n_\mathrm{i}/N$, then we have

$$\langle f_\mathrm{i}\rangle = p_\mathrm{i} \quad \text{and} \quad \langle(\delta f_\mathrm{i})^2\rangle = \frac{p_\mathrm{i}}{N}. \tag{A395}$$

But if we identify frequencies as our best estimate of probabilities (and we'll see below in what sense this familiar identification is correct), we can construct a 'naive' estimate of the entropy,

$$S_\mathrm{naive} = -\sum_{\mathrm{i}=1}^{K} f_\mathrm{i}\log_2 f_\mathrm{i}. \tag{A396}$$

Since the frequencies are close to the true probabilities when the number of samples is large, we can do a Taylor expansion around the point $f_\mathrm{i} = \langle f_\mathrm{i}\rangle = p_\mathrm{i}$:

$$\begin{aligned} S_\mathrm{naive} &= -\sum_{\mathrm{i}=1}^{K} f_\mathrm{i}\log_2 f_\mathrm{i} \\ &= -\sum_{\mathrm{i}=1}^{K}(p_\mathrm{i}+\delta f_\mathrm{i})\log_2(p_\mathrm{i}+\delta f_\mathrm{i}) \qquad \text{(A397)} \\ &= -\sum_{\mathrm{i}=1}^{K} p_\mathrm{i}\log_2 p_\mathrm{i} - \sum_{\mathrm{i}=1}^{K}\left[\log_2 p_\mathrm{i} + \frac{1}{\ln 2}\right]\delta f_\mathrm{i} - \frac{1}{2}\sum_{\mathrm{i}=1}^{K}\left[\frac{1}{(\ln 2)p_\mathrm{i}}\right](\delta f_\mathrm{i})^2 + \cdots . \qquad \text{(A398)} \end{aligned}$$

The first term in the series is the true entropy. The second term is a random error which averages to zero. The third term, however, has a nonzero mean, since it depends on the square of the fluctuations δf_i. Thus when we compute the average of our naive entropy estimate we find

$$\begin{aligned} \langle S_\mathrm{naive}\rangle &= S_\mathrm{true} - \frac{1}{2\ln 2}\sum_{\mathrm{i}=1}^{K}\frac{\langle(\delta f_\mathrm{i})^2\rangle}{p_\mathrm{i}} + \cdots \qquad \text{(A399)} \\ &= S_\mathrm{true} - \frac{1}{2\ln 2}\sum_{\mathrm{i}=1}^{K}\frac{p_\mathrm{i}}{Np_\mathrm{i}} + \cdots \qquad \text{(A400)} \\ &= S_\mathrm{true} - \frac{K}{2\ln 2N} + \cdots . \qquad \text{(A401)} \end{aligned}$$

Thus, no matter what the underlying true distribution, identifying frequencies with probabilities leads to a *systematic* (not random!) underestimate of the entropy, and the size of this systematic error is proportional to the number of accessible states (K) and inversely proportional to the number of samples (N).

The fact that the systematic errors have a very definite structure suggests that we should be able to correct them. Let us see what happens to our entropy estimates in the "experiment" of Fig 175 as we change the number of samples N. More precisely, suppose we have only the $N = 100$ samples, but we choose $n < 100$ points out of these samples, and estimate the entropy based only on this more limited data. Equation (A401) suggests that if we plot our entropy estimate vs. $1/n$, we should see a straight line; a higher order version of the same calculation shows that there are quadratic corrections. Indeed, as shown in Fig 176, this works. It is important to note that, for all the accessible range of sample sizes, the entropy estimate is smaller than the true entropy, and this error is larger than our best estimate of the error bar; this really is dangerous. On the other hand, once we recognize the systematic dependence of the entropy estimate on the number of samples, we can extrapolate to recover an estimate that is correct within error bars. What we have seen here about entropy is also true about information, which is a difference between entropies.

It is also important to show that this extrapolation procedure works also for real data, not just for the idealized case where we choose samples independently out of a known distribution. Decide what examples to use. One from neurons, one from genes, one from sequences?

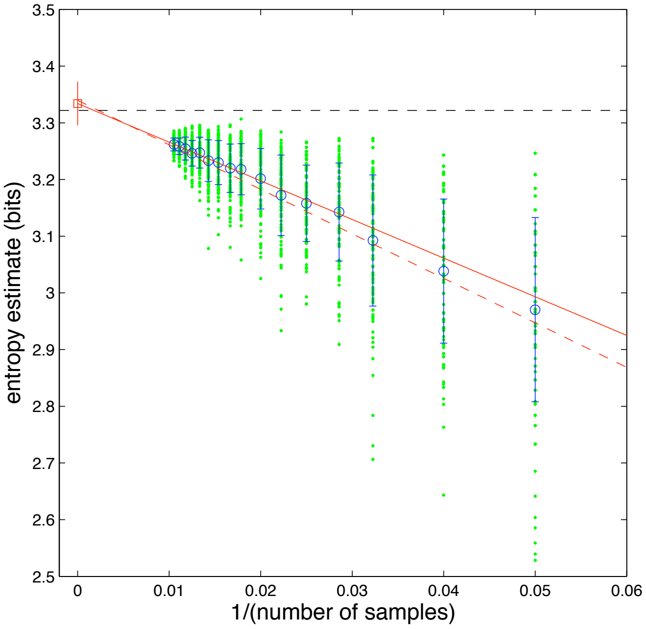

FIG. 176 Entropy vs. number of samples. Starting with $N = 100$ samples, as in Fig 175, we draw smaller numbers of samples at random, compute the entropy, and search for the systematic behavior predicted in Eq (A401). Green points are from different subsamplings, blue circles show means and standard deviations. Red line is a linear fit for $n > N/2$, and red dashed line is a quadratic fit to all of the data shown. Red square is the extrapolation, with an error bar $\sqrt{2}$ smaller than the standard deviation found empirically at $n = N/2$, and the dashed black line is $S_{\rm true} = \log_2(10)$.

So far one figure from neurons, Fig 177.

One might worry that entropy estimates based on extrapolations are a bit heuristic. If we can really convince ourselves that we see a clean linear dependence on $1/N$, things are likely to be fine, but this leaves room for considerable murkiness. Also, since the expansion of the entropy estimate in powers of $1/N$ obviously is not fully convergent, there is always the problem of choosing the regime over which the asymptotic behavior is observed, a widespread problem in fitting to such asymptotic series. While for many purposes these problems can be dismissed, it would be nice to do better. It also is an interesting mathematical challenge to ask if we can estimate the entropy of a probability distribution even when the number of samples we have seen is small, perhaps even smaller than the number of possible states for the system.

Whenever we do a Monte Carlo simulation of a physical system in thermal equilibrium, we are in the "undersampled" limit, where the number of samples we collect must be much smaller than the number of possible states. Usually if we want to estimate entropy from Monte Carlo, we use the identity which relates entropy to an integral of the heat capacity, since heat capacity is related to energy fluctuations and these are easy to compute at each temperature. Of course, if you just have samples of the state of the system, and don't actually know the Hamiltonian, you can't compute the energy and so this doesn't work. Ma suggested another approach, asking how often the system revisits the same state. In the simple case (relevant for the microcanonical ensemble) where all K possible states are equally likely, the probability that two independent samples are in the same state is $1/K$. But if we have N samples, we have $\sim N^2$ pairs that we can test. Thus we can get a good estimate of the probability of occupying the same state once we observe $N \sim \sqrt{K}$ independent samples, far less than the number of states. As an illustration, Fig 178 shows the frequency of coincidences when we draw N samples from a uniform distribution with $K = 100$ states.

We recall the classic problem of how many people need to be in the room before there is a good chance of two people have the same birthday. The answer is not 365, but more nearly $\sqrt{365}$. Put another way, if we didn't know the length of the year, we could estimate this by polling people about their birthdays, and keeping track of coincidences. Long before we have sampled all possible birthdays, Fig 178 shows us that our estimate of this coincidence probability will stabilize—which birthdays are represented will vary from sample to sample, but the fraction of coincidences will vary much less.

In these simple examples, the probability distribution is uniform, and so the entropy is just the log of the num-

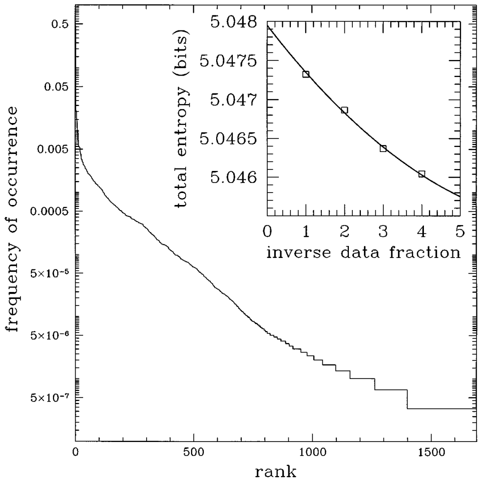

FIG. 177 Entropy extrapolation with real neural data, from Strong et al (1998a). From the experiment on fly motion–sensitive neurons discussed in Figs 132 and 133, we look at 10–letter words with time resolution $\Delta\tau = 3\,{\rm ms}$. The main figure shows the "Zipf plot" of frequency vs. rank from the full data set. Note that since there are sometimes (but rarely) two spike in one 3 ms bin, there are more than 1024 words. The inset shows the estimated entropy as a function of the (inverse) fraction of the full data set used. The line through the data is from Eq (A401).

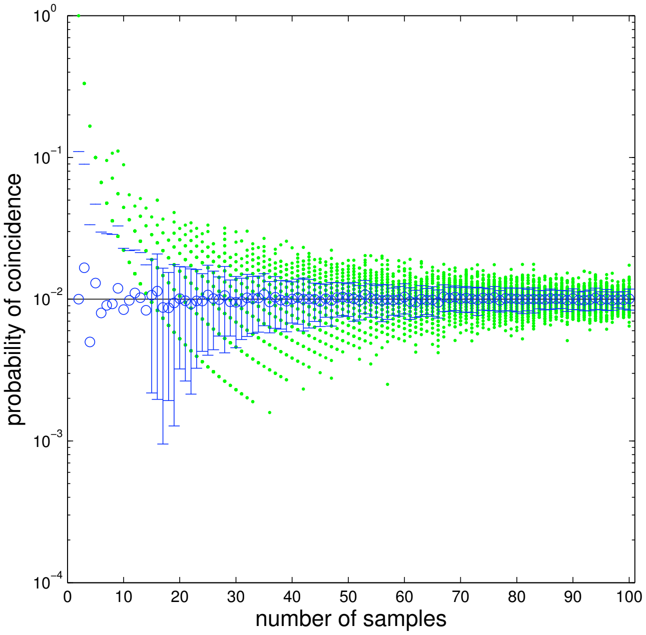

FIG. 178 Estimating coincidence probability. Samples are drawn from a distribution that is uniform over $K = 100$ possible states. Green dots show examples, blue circles the mean and standard deviation across many draws of N samples, and black dashed line is the exact answer. We see that the estimate is quite good even when $N \sim \sqrt{K} \ll K$.

ber of possible states, and this in turn is inversely proportional to the probability of a coincidence. So, being able to estimate this probability is equivalent to being able to estimate the entropy. Thus we should be able to generate reliable entropy estimates even in the undersampled regime, just by counting coincidences. This is a beautiful idea. The challenge is to generalize this idea to non–uniform distributions.

A better understanding of the entropy estimation problem has come through a Bayesian approach. Rather than identifying frequencies with probabilities, we imagine that the distribution itself is drawn from a distribution. To be formal, let the possible states of the system be $\mathrm{i} = 1, 2, \cdots, K$, and let the probability distribution over these states be $p_1, p_2, \cdots, p_K \equiv \mathbf{p}$. This distribution itself is drawn from some distribution function $\mathcal{P}(\mathbf{p})$. The distribution has to be normalized, but it is tempting to think that, other than normalization, all distributons should be equally likely, so that

$$\mathcal{P}(\mathbf{p}) = \frac{1}{Z}\delta\left(\sum_{\mathrm{i}=1}^{K} p_\mathrm{i} - 1\right). \tag{A402}$$

If we observe n_1 samples in the first state, n_2 samples in the second state, and so on, then the probability of this occurring assuming some distribution $\mathbf{p}$ is

$$P(\{n_\mathrm{i}\}|\mathbf{p}) \propto \prod_{\mathrm{i}=1}^{K} p_\mathrm{i}^{n_\mathrm{i}}, \tag{A403}$$

and so by Bayes' rule we have

$$\mathcal{P}(\mathbf{p}|\{n_\mathrm{i}\}) = \frac{P(\{n_\mathrm{i}\}|\mathbf{p})\mathcal{P}(\mathbf{p})}{P(\{n_\mathrm{i}\})} \tag{A404}$$

$$\propto \frac{1}{Z}\left(\prod_{\mathrm{i}=1}^{K} p_\mathrm{i}^{n_\mathrm{i}}\right)\delta\left(\sum_{\mathrm{i}=1}^{K} p_\mathrm{i} - 1\right). \tag{A405}$$

If we want to compute our best estimate of the distribution, we have to do the integral

$$\hat{p}_\mathrm{i} = \frac{1}{\mathcal{Z}}\int d^K\mathbf{p}\, p_\mathrm{i}^{n_\mathrm{i}+1}\left(\prod_{\mathrm{j}\neq\mathrm{i}} p_\mathrm{j}^{n_\mathrm{j}}\right)\delta\left(\sum_{\mathrm{j}=1}^{K} p_\mathrm{j} - 1\right), \tag{A406}$$

where the normalization $\mathcal{Z}$ is given by

$$\mathcal{Z} = \int d^K\mathbf{p}\left(\prod_{\mathrm{j}=1}^{K} p_\mathrm{j}^{n_\mathrm{j}}\right)\delta\left(\sum_{\mathrm{j}=1}^{K} p_\mathrm{j} - 1\right). \tag{A407}$$

To make progress we introduce the Fourier representation of the delta function, so that, for example,

$$\mathcal{Z} = \int d^K\mathbf{p}\left(\prod_{\mathrm{j}=1}^{K} p_\mathrm{j}^{n_\mathrm{j}}\right)\int\frac{d\lambda}{2\pi}\exp\left(+i\lambda\sum_{\mathrm{j}=1}^{K} p_\mathrm{j} - i\lambda\right) \tag{A408}$$

$$= \int\frac{d\lambda}{2\pi}e^{-i\lambda}\prod_{\mathrm{j}=1}^{K}\int dp_\mathrm{j}\, p_\mathrm{j}^{n_\mathrm{j}}e^{i\lambda p_\mathrm{j}}. \tag{A409}$$

Since we have the delta function, we are free to let the integrals over p_j run from 0 to ∞; the delta function will enforce the constraint that $p_\mathrm{i} \leq 1$ for all i. Then the key ingredient of the calculation, then, is the integral

$$f(n;\lambda) = \int_0^\infty dp\, p^n e^{i\lambda p}. \tag{A410}$$

At the end of our calculation we will have to integrate over λ. Let's assume that we will be able to deform the contour of this integral into the complex λ plane in such a way that the p integral in Eq (A410) is well behaved. Then we can write

$$f(n;\lambda) = \int_0^\infty dp\, p^n e^{i\lambda p} \tag{A411}$$

$$= \int_0^\infty dp\, p^n e^{-(-i\lambda)p} = \frac{n!}{(-i\lambda)^{n+1}}. \tag{A412}$$

Putting these pieces together, we have

$$\mathcal{Z} = \int \frac{d\lambda}{2\pi} e^{-i\lambda} \prod_{\mathrm{j}=1}^{K} \frac{n_\mathrm{j}!}{(-i\lambda)^{n_\mathrm{j}+1}} \tag{A413}$$

$$= \left(\prod_{\mathrm{j}=1}^{K} n_\mathrm{j}!\right) \int \frac{d\lambda}{2\pi} \frac{e^{-i\lambda}}{(-i\lambda)^{\sum_{\mathrm{j}=1}^{K}(n_\mathrm{j}+1)}} \tag{A414}$$

$$= \left(\prod_{\mathrm{j}=1}^{K} n_\mathrm{j}!\right) \int \frac{d\lambda}{2\pi} \frac{e^{-i\lambda}}{(-i\lambda)^{N+K}}, \tag{A415}$$

where $N = \sum_\mathrm{j} n_\mathrm{j}$ is the total number of samples, and as before K is the number of possible states. A similar argument gives

$$\hat{p}_\mathrm{i} = \frac{1}{\mathcal{Z}}(n_\mathrm{i}+1)! \left(\prod_{\mathrm{j}\neq\mathrm{i}} n_\mathrm{j}!\right) \int \frac{d\lambda}{2\pi} \frac{e^{-i\lambda}}{(-i\lambda)^{N+K+1}} \tag{A416}$$

$$= \frac{(n_\mathrm{i}+1)! \left(\prod_{\mathrm{j}\neq\mathrm{i}} n_\mathrm{j}!\right)}{\prod_{\mathrm{j}=1}^{K} n_\mathrm{j}!} \times \frac{\int \frac{d\lambda}{2\pi} \frac{e^{-i\lambda}}{(-i\lambda)^{N+K+1}}}{\int \frac{d\lambda}{2\pi} \frac{e^{-i\lambda}}{(-i\lambda)^{N+K}}} \tag{A417}$$

$$= (n_\mathrm{i}+1) \frac{\int \frac{d\lambda}{2\pi} \frac{e^{-i\lambda}}{(-i\lambda)^{N+K+1}}}{\int \frac{d\lambda}{2\pi} \frac{e^{-i\lambda}}{(-i\lambda)^{N+K}}}. \tag{A418}$$

Thus, $\hat{p}_\mathrm{i} \propto n_\mathrm{i} + 1$, so to get the normalization right we must have

$$\hat{p}_\mathrm{i} = \frac{n_\mathrm{i}+1}{N+K}. \tag{A419}$$

This should be contrasted with the naive estimate of probabilities based on counting frequencies, $\hat{p}_\mathrm{i} = n_\mathrm{i}/N$. The Bayesian estimate, with a 'flat' prior on the space of distributions, is equivalent to the naive approach but with one extra count in every bin. This estimate never predicts probability zero, even in states never observed to occur, and is in some sense 'smoother' than the frequencies. The trick of adding such pseudocounts to the data goes back, it seems, to Laplace, although I don't think he had the full Bayesian justification.

Problem 196: Normalization. Derive Eq (A419) directly by doing the integrals in Eq (A418).

What does this have to do with entropy estimation? Somewhat heuristically, it has been suggested that by using different numbers of pseudocounts one can improve the quality of entropy estimation. More deeply, I think, the Bayesian estimate gives us a very different view of *why* we make systematic errors when we try to compute entropies from data. Recall that when we use the naive identification of frequencies with probabilities, we underestimate the entropy, as in Eq (A401). It is tempting to think that we are underestimating the entropy simply because, in a finite sample, we have not seen all the possibilities. With the Bayesian approach and a flat prior, however, the probability distributions that we estimate are smoother than the true distribution, and correspondingly we expect that the entropy will be overestimated. In fact this is true, but the problem really is more serious than this.

Suppose that we don't yet have any data. Then all we know is that the probability distribution $\mathbf{p}$ will be chosen out of the distribution $\mathcal{P}(\mathbf{p})$. This seems innocuous, since this distribution is flat and hence presumably unbiased. But we can calculate the average entropy in this distribution,

$$\langle S\rangle_{\mathrm{prior}} \equiv \int d^K\mathbf{p} \left(-\sum_{\mathrm{i}=1}^{K} p_\mathrm{i} \log_2 p_\mathrm{i}\right) \mathcal{P}(\mathbf{p}), \tag{A420}$$

using the same tricks that we used above, and we find

$$\langle S\rangle_{\mathrm{prior}} = \psi_0(K+1) - \psi_0(1), \tag{A421}$$

where $\psi_0(x)$ is a polygamma function,

$$\psi_m(x) = \left(\frac{d}{dx}\right)^{m+1} \Gamma(x). \tag{A422}$$

The details of the special functions are not so important. What is important is that, when the number of states K is large,

$$\langle S\rangle_{\mathrm{prior}} = \log_2 K - \mathcal{O}(1). \tag{A423}$$

Thus, although we are choosing distributions from a flat prior, the entropies of these distributions are biased toward the maximum possible value. This bias is actually very strong. The entropy is the average of many terms, and although these terms can't be completely independent (the probabilities must sum to one), one might expect the central limit theorem to apply here, in which case the fluctuations in the entropy will be $\sigma_S \sim 1/\sqrt{K}$, which for large K is very small indeed. What this means is that the distributions chosen out of $\mathcal{P}(\mathbf{p})$ are overwhelmingly biased toward having nearly maximal entropy. While the prior on the distributions is flat, the prior on entropies is narrowly concentrated around an average entropy which, for large K, is almost $\log_2 K$.

Problem 197: Entropies in a flat prior. Derive the mean and standard deviation of the entropy in the flat prior, $\mathcal{P}(\mathbf{p})$ from Eq (A402). Verify Eq (A423).

Just to make the problem clear, suppose that our system has only two states, as with heads and tails for a coin. Let the probability of heads be q, so that the entropy is

$$S(q) = -q\log_2(q) - (1-q)\log_2(1-q). \qquad \text{(A424)}$$

If we assume that q is chosen from some distribution $\mathcal{P}(q)$, then the distribution of entropies can be found from

$$P(S)dS = \mathcal{P}(q)dq \qquad \text{(A425)}$$

Since $dS/dq = 0$ at the point where $S = 1\,\text{bit}$, the distribution $P(S)$ must be singular there unless the prior on q itself has a compensating singularity. Thus, a prior which is flat in q is strongly biased in S. The situation is even worse for systems with many states, because of phase space considerations: if we want to have low a low entropy distribution, then many of the $p_{\rm i}$ must be confined to very small values, and this means that the volume in $\mathbf{p}$ space associated with low entropy is small. While only one distribution has precisely the maximum entropy, there are many distributions that are close.

Problem 198: A flat prior on S. Show that, for the problem of coin flips, having a flat prior on the entropy S is equivalent to a prior

$$\mathcal{P}(q) = \left|\log_2\left(\frac{q}{1-q}\right)\right|. \qquad \text{(A426)}$$

If we flip a coin N times and observe n heads, then Bayes' rule tell us that

$$\mathcal{P}_N(q|n) \propto \mathcal{P}(q) q^n (1-q)^{N-n}, \qquad \text{(A427)}$$

and we can use this to estimate the entropy

$$\hat{S}(n,N) = \int_0^1 dq\,\mathcal{P}_N(q|n)\left[-q\log_2(q) - (1-q)\log_2(1-q)\right]. \qquad \text{(A428)}$$

(a.) For $N = 10$, plot $\hat{S}(n,N)$ vs. n. Compare your results with the naive estimate,

$$S_{\rm naive}(n,N) = -\frac{n}{N}\log_2\left(\frac{n}{N}\right) - \left(1-\frac{n}{N}\right)\log_2\left(1-\frac{n}{N}\right). \qquad \text{(A429)}$$

(b). Suppose that you are actually flipping a coin in which the probability of heads is $q_{\rm true} \neq 1/2$. Simulate N such flips, and use your results to estimate the entropy according to both Eq's (A428) and (A429). How do these estimators evolve as a function of N? Hints: Remember that we have seen the results for the naive estimator already, and that since this is a small system (only two states) the interesting behavior is at smaller N.

[Add figure on entropy estimation for binary variables, with flat priors on entropy or probability.] All of this suggests that we could do a much better job of entropy estimation in a Bayesian framework where the $\mathcal{P}(\mathbf{p})$ is chosen to be flat is S. I don't know of anyone who has given a complete solution to this problem. A partial solution has been proposed by noticing that there is a well known generalization of the flat prior, the Dirichlet family of priors

$$\mathcal{P}_\beta(\mathbf{p}) = \frac{1}{Z(\beta)}\left(\prod_{\rm i-1}^{K} p_{\rm i}^{\beta-1}\right)\delta\left(\sum_{\rm i=1}^{K} p_{\rm i} - 1\right). \qquad \text{(A430)}$$

Evidently the flat prior corresponds to $\beta = 1$, and this is biased toward large entropies, as we have seen. As β gets smaller, the average entropy $\bar{S}(\beta)$ of a distribution drawn out of $\mathcal{P}_\beta(\mathbf{p})$ gets smaller, but for each value of β the distribution of entropies remains quite narrow. This suggests that if we form the prior

$$\mathcal{P}(\mathbf{p}) = \int_0^1 d\beta \left|\frac{d\bar{S}(\beta)}{d\beta}\right|^{-1}\mathcal{P}_\beta(\mathbf{p}), \qquad \text{(A431)}$$

it will be approximately flat in entropy. This seems to work, although it is computationally intensive. As far as I know it gives the best results of any estimation procedure so far in, for example, the analysis of neural spike trains. If we dig into the integrals that define the entropy estimate, it turns out that the key pieces of data are coincidences, in which more than one sample falls into the same bin, and in this sense we seem to have found a generalization of Ma's ideas to non–uniform distributions.

Problem 199: One more problem about entropy estimation. [make up one more?]

Again it is important to ask whether these ideas actually work with real data. In experiments on the motion–sensitive neuron H1 in the fly visual system, we can in many cases collect enough data to sample the underlying distributions of neural responses, so we have ground truth. At the same time, we can look only at a small fraction of these data and ask how well our estimation procedure works. An example is shown in Fig 179.

Is there more to say? Or need more details in things already said?

The idea that naive counting leads to systematic errors in entropy estimation goes back, at least, to Miller (1955). The importance of this for the analysis of information transmission in neurons was emphasized by Treves and Panzeri (1995), who also brought more sophistication to the calculation of the series expansion that we have started here. Shortly after this, Strong et al (1998a) showed how these extrapolation methods could be used to estimate entropy and information in neural responses to complex, dynamic sensory inputs. An important technical point is that Strong et al took seriously the $1/N$ behavior of the entropy estimate, but didn't use an analytic calculation of the slope of $S_{\rm est}$ vs $1/N$; the reason is

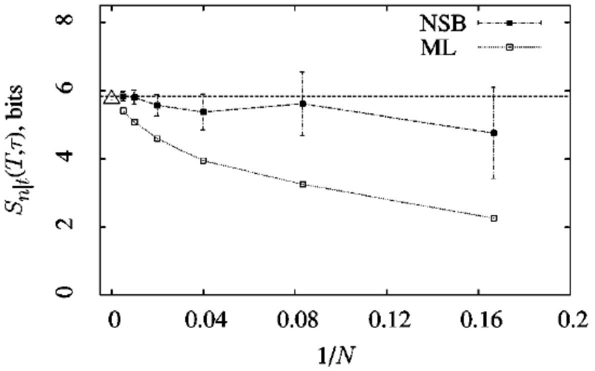

FIG. 179 Estimating entropies at one slice of time in the neural response to naturalistic stimuli, from Nemenman et al (2004). Neural responses are discretized with $\Delta\tau = 2\,\text{ms}$ resolution, and we look at 8–letter words. The stimulus is motion outdoors, and the motion is repeated many times; here we focus on the distribution of responses at one moment relative to this repeat, for which we can collect up to 196 samples from the repetitions. The open symbols show the "naive" or maximum likelihood estimate in which we identify the observed frequencies with probabilities and plug in to the computation of entropy. As expected, this estimate has a significant dependence on the number of samples, but extrapolates smoothly according to Eq (A401). In contrast, the NSB estimator based on the prior in Eq (A431) remains constant within error bars, always agreeing with the extrapolation.

that some seemingly possible neural responses are expected to have probability zero, because there is a hard core repulsion ("refractoriness") between spikes, but we don't know in advance exactly how big this effect will be. As a result, the actual number of possible states K is uncertain, and in addition it is not true that all the samples collected in the experiment will be independent. Both these effects leave the $1/N$ behavior intact, but change the slope.

Miller 1955: Note on the bias of information estimates. GA Miller, in *Information Theory in Psychology: Problems and Methods II–B*, H Quastler, ed pp. 95–100 (Free Press, Glencoe IL, 1955).

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).

Treves & Panzeri 1995: The upward bias in measures of information derived from limited data samples. A Treves & S Panzeri, *Neural Comp* **7,** 399–407 (1995).

The idea of estimation entropy by counting coincidences is presented in Ma (1981), although it has precursors in Serber (1973). The attempts to build a 'flat prior' on the entropy are described in a series of papers by Nemenman (2002) and co–workers (Nemenman et al 2002). Moments of the entropy distribution in Dirichlet priors were calculated by Wolpert & Wolf (1995), so you can see where the polygamma functions come from. Within the family of Dirichlet priors, we have noted that the flat prior on distributions ($\beta = 1$) was discussed by Laplace (1814) as the idea of starting with one pseudocount in every bin; Jeffreys (1946), and later Krichevskii & Trofimov (1981) proposed using half a pseudocount ($\beta = 1/2$), while Schurmann & Grassberger (1996) suggested that entropy estimates could be improved by adjusting the number of pseudocounts to the number of bins, $\beta = 1/K$. The procedure developed by Nemenman et al integrates over all β, allowing a dominant β to emerge that is matched to the structure of the data and to the number of samples, as well as to K. The example in Fig 179 is from Nemenman et al (2004).

Jeffreys 1946: An invariant form for the prior probability in estimation problems. H Jeffreys, *Proc R Soc Lond Ser A* **186,** 453–461 (1946).

Krichevskii & Trofimov 1981: The performance of universal encoding. R Krichevskii & V Trofimov, *IEEE Trans Inf Thy* **27,** 199–207 (1981)

Laplace 1814: *Essai philosophique sur les probabilités* (Courcier, Paris, 1814). *A Philosophical Essay of Probabilities*, translated by F Truscott & F Emory (Dover, New York, 1951).

Ma 1981: Calculation of entropy from data of motion. S–K Ma, *J Stat Phys* **26,** 221–240 (1981).

Nemenman 2002: Inference of entropies of discrete random variables with unknown cardinalities. I Nemenman, arXiv:physics/0207009 (2002).

Nemenman et al 2002: Entropy and inference, revisited. I Nemenman, F Shafee & W Bialek, in *Advances in Neural Information Processing 14,* TG Dietterich, S Becker & Z Ghahramani, eds, pp 471–478 (MIT Press, Cambridge, 2002); arXiv:physics/0108025 (2001).

Nemenman et al 2004: Entropy and information in neural spike trains: Progress on the sampling problem. I Nemenman, W Bialek & R de Ruyter van Steveninck, *Phys Rev E* **69,** 056111 (2004); arXiv:physics/0306063 (2003).

Serber 1973: *Estimation of Animal Abundance and Related Parameters.* GAF Serber (Griffin, London, 1973).

Schurmann & Grassberger 1996: Entropy estimation of symbol sequences. T Schurmann & P Grassberger, *Chaos* **6,** 414–427 (1996).

Victor 2002: Binless strategies for estimation of information from neural data. J Victor *Phys Rev E* **66,** 051903 (2002).

Wolpert & Wolf 1995: Estimating functions of probability distributions from a finite set of samples. DH Wolpert & DR Wolf, *Phys Rev E* **52,** 6841–6854 (1995).

The specific problem of estimating entropy in neural responses has the added feature that spikes are discrete, but they can occur at any time, so there is a question of whether we should view the whole problem as discrete (with bins along the time axis) or continuous (with a metric along the time axis); the discussion here has focused on discrete problems. For metric space approaches to spike trains, see Victor (2002). For a general overview of the entropy estimation problem, with particular attention to the challenges posed by neural data, see Paninski (2003), who tries to make many of the heuristic arguments in the field more rigorous. One can also view entropy estimation as a problem in computational complexity—how many samples, and hence how many computational steps, do we need in order to approximate the entropy to some level of accuracy? For an approach in this spirit, see Batu et al (2002). [One more recent paper comparing the effectiveness of different approaches]

Batu et al 2002: The complexity of approximating the entropy. T Batu, S Dasgupta, R Kumar & R Rubinfeld, in *Proc 34th Symp Theory of Computing (STOC)*, pp 678–687 (2002).

Paninski 2003: Estimation of entropy and mutual information. L Paninski, *Neural Comp* **15,** 1191–1253 (2003).

Victor 2002: Binless strategies for estimation of information from neural data. J Victor *Phys Rev E* **66,** 051903 (2002).