These notes will take you through our calculation of the optimal stimulus transform, $\Phi(m)$, of a stimulus message, m, that maximizes the response, r, entropy, $S(r)$, in a Gaussian channel. We learned at the end of class that some neurons in the fly visual system actually obey this optimal relationship between the stimulus distribution, $P(m)$, and the response, $r(m)$.

Preliminaries: We started class by reviewing some results from information theory. We asserted a few facts that are easy enough to prove on your own, like the overly-fancily-labeled Bayes' Rule, which follows very simply from the fact that $P(A, B) = P(B, A)$, and an identity on the joint probability, $P(A, B) = P(A|B)P(B)$:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

Next, we defined the average uncertainty about some variable X that takes values x_i. The formula for the average uncertainty about X's value is its entropy,

$$S(X) = -\sum_i P(x_i) \log_2 P(x_i). \tag{2}$$

We noted that the choice of the base for the logarithm really doesn't matter, but most info theory works choose base 2, so your entropies and information quantities are in units of 'bits'. Using the natural logarithm gives you 'nats' or 'nits'. Using base 10 yields 'dats' or 'dits'. For obvious reasons, folks never use 'bats' to describe base 2 units of information.

The mutual information between two variables X and Y can be understood as the generalized correlation between these variables (to all orders), or perhaps more intuitively, the reduction in uncertainty about one variable provided by knowing the other. The mutual information has the form

$$I(X; Y) = S(X) - \langle S(X|Y) \rangle_y. \tag{3}$$

It's a positive quantity (see separate notes that prove this using Gibb's inequality; you can also prove this using Jensen's inequality), it's symmetric, and it's a special case of the Kullback-Leibler divergence (D_{KL}), between two distributions. You can see this last fact most easily by rewriting the mutual information in terms of the joint distribution of X and Y to yield

$$I(X; Y) = \sum_{i,j} P(x_i, y_j) \log_2 \left(\frac{P(x_i, y_j)}{P(x_i)P(y_j)} \right). \tag{4}$$

The D_{KL} has the form

$$D_{\text{KL}}(P||Q) = \sum_i P(x_i) \log \frac{P(x_i)}{Q(x_i)}. \tag{5}$$

The mutual information is a particular instance of the D_{KL}, where the two distributions we are comparing are the joint distribution, $P(X, Y)$, and the product of the marginals, $P(X)P(Y)$. Clearly, if and only if X and Y are independent, these two distributions are eqivalent.

Information in a Gaussian channel: We are sending messages, m, through a response function, r, that makes some transform on the stimulus, Φ, and has noise such that

$$r = f(m) = \Phi(m) + z, \tag{6}$$

where z is independent, Gaussian noise with zero mean and variance σ_z^2. Now let's write down the mutual information between the response and the stimulus message.

$$I(R;M) = S(R) - \langle S(R|m)\rangle_m \tag{7}$$
$$= S(R) - \langle S(\Phi(m) + z|m)\rangle_m \tag{8}$$

. . . at this step we invoked the translation invariance of the entropy, because $\Phi(m)$ given a particular m is a constant. . .

$$= S(R) - \langle S(Z|m)\rangle_m \tag{9}$$
$$= S(R) - S(Z) \tag{10}$$

In the last step, we use the fact that the noise is independent of the stimulus message, m, so we simply recover the noise entropy. If we have a fixed noise source, as we often do in biological systems, then $S(Z)$ is given and the only thing we can do to maximize $I(R;M)$ is maximize $S(R)$. So now that we have that result, let's derive the distribution that maximizes $S(R)$.

We want to maximize the entropy while ensuring that the probability distribution remains normalized, so we'll use the method of Lagrange multipliers here, though there are other ways to solve for the optimal $P(r)$:

$$\mathcal{L} = -\sum_i P(r_i)\log P(r_i) - \lambda\left(1 - \sum_i P(r_i)\right) \tag{11}$$
$$\frac{\delta\mathcal{L}}{\delta P(r_k)} = -\frac{P(r_k)}{P(r_k)} - \log P(r_k) + \lambda = 0 \tag{12}$$
$$P(r_k) = e^{1-\lambda} \tag{13}$$

where we note that the optimal $P(r)$ doesn't have a k dependence, i.e. is constant. To be clear, the δ symbol here is a variational derivative, also called a functional derivative. We are trying to compute how $\mathcal{L}$ changes when we vary the parameters $P(r_k)$. These are all functions of r_k, so one should be careful with notation. Here, our functional, $\mathcal{L}(P(r))$, is a sum (or for the continuous case, an integral) of some function of $P(r)$. What we need to do is expand our functional in powers of δP. The linear term is defined as our functional derivative. We do not encounter any dependence of our summands (integrands) on derivatives of $P(r)$, so our functional derivative is, in this case, equal to the partial derivative, $\partial\mathcal{L}/\partial P(r)$. Check out the wiki page on the Euler-Lagrange equation if you'd like to know more. Practically speaking, you can think of this like a continuous version of a partial derivative, and mechanically, you will take derivatives in the usual way.

Solving for λ we have that

$$\sum_i e^{1-\lambda} = 1 \tag{14}$$

$$Ne^{1-\lambda} = 1 \tag{15}$$

$$e^{1-\lambda} = \frac{1}{N} \tag{16}$$

$$\implies P(r_k) = \frac{1}{N}, \quad \forall r_k \tag{17}$$

where N is the number of response levels. If we imagine that we have binned our response states between 0 and some maximal response r_{max} using a bin size Δr, then we have that

$$P(r_k) = \frac{\Delta r}{r_{\text{max}}}, \quad \forall r_k \tag{18}$$

We have shown that the distribution that maximizes the entropy for a discrete, normalized probability distribution is the uniform distribution. You can show that maximizing entropy while constraining the mean rate of response, and confining responses to be greater than or equal to zero leads to an exponential distribution of responses. Maximizing entropy while constraining the mean and variance (and relaxing the positivity constraint) leads to a Gaussian distribution.

Keeping with our uniform maximum response entropy distribution, we now want to derive the relationship between the response r and the stimulus. In particular, we want to derive the form of $\Phi(m)$ that gives us a uniform distribution of $P(r)$, for some distribution of the stimulus messages, $P(m)$.

We start with a continuity condition on the probability density functions for r and m, taking the limit of small Δr and Δm. For a small change in the message, Δm, the probability that the message falls between m and $m + \Delta m$ is simply $p(m)\Delta m$. This creates a response, r, that falls between $\Phi(m)$ and $\Phi(m + \Delta m)$. The probability that the response falls in this range is $p(r)\Delta r$, with $\Delta r = \Phi(m + \Delta m) - \Phi(m)$ and $p(r) = p(\Phi(m))$. Taking this step by step:

$$p(r)\Delta r = p(m)\Delta m \tag{19}$$

where $p(r)$ is now the probability *density* function for r, $p(r) = P(r)/\Delta r$.

$$p(r)\Delta r = p(m)\Delta m \tag{20}$$

$$p(r)(\Phi(m + \Delta m) - \Phi(m)) = p(m)\Delta m \tag{21}$$

$$\tag{22}$$

Now, let's fill in the $p(r)$ that maximizes the response entropy:

$$\frac{1}{r_{\text{max}}}(\Phi(m + \Delta m) - \Phi(m)) = p(m)\Delta m \tag{23}$$

$$\frac{(\Phi(m + \Delta m) - \Phi(m))}{\Delta m} = r_{\text{max}} p(m) \tag{24}$$

If $\Phi(m + \Delta m) \geq \Phi(m)$, i.e. if Φ is monotonic in m, and if $\Delta m \to 0$, we can rewrite the left hand side as the derivative of Φ:

$$\frac{d\Phi}{dm} = r_{\text{max}} p(m) \tag{25}$$

Integrating both sides yields

$$\Phi(m) = r_{\text{max}} \int_{m_{\text{min}}}^{m} p(m') dm' \tag{26}$$

This holds for monotonic $\Phi(m)$ and small Δm. To obtain a uniform response distribution, which we have shown maximizes the response entropy and thereby the mutual information between the stimulus, m, and the response, $r(m)$, the response should be the CDF of the stimulus distribution up to the stimulus value, m. Simon Laughlin was the first to show that this response form is actually observed in biological systems, in a seminal paper on large monopolar cells in the blowfly retina in 1981.