M2 Biology

TD9: Coding I

Elie Oriol

`https://github.com/Elieoriol/2021_UlmM2_ThNeuro/tree/master/TD9`

This supplementary tutorial tackles some topics of coding in Neuroscience, namely mutual information, Fisher information, and Bayesian inference.

1 Mutual Information

1.1 Characterizing the distribution of a discrete random variable

Suppose X is a random variable which can take the discrete values $\{X_1, ..., X_n\}$ with probability $p(X_i)$. The entropy of the distribution is defined as:

$$H(X) = -\sum_{i=1}^{n} p(X_i) \log(p(X_i))$$

This is also called the "information" of the distribution, and we shall see why. Suppose that the X_i correspond to the possible colors of a ball picked randomly from an urn. I observe the color of the ball, and you can ask me questions which I can only answer with yes or no. Calculate the entropy and the average number of questions needed to find out the color of the ball in the following cases:

1. If there is only one possible color.

   In that case you don't even need to ask any question, you already know the color of the ball: 0 questions. The entropy is $H = -1 \log(1) = 0$.

2. If there are two possible colors.

   One question is sufficient, for instance "Is it red?" if the possible colors are blue and red. The entropy is $H = -p \log(p) - (1-p) \log(1-p)$*, which is equal to* 0 *for* $p = 0$ *or* 1*, and maximal at* $p = 0.5$ *where it is equal to* $\log(2)$.

3. If half the balls are red, one fourth are green and one fourth are blue.

   A set of questions could start with "Is it red?" $\rightarrow$ 1/2 *chance that the answer is yes, or* 1/2 *chance that the answer is no, in which case another question is needed to distinguish green and blue. The entropy is:*

$$H = -\frac{1}{2}\log(\frac{1}{2}) - \frac{1}{4}\log(\frac{1}{4}) - \frac{1}{4}\log(\frac{1}{4}) = \frac{1}{2}\log(2) + \frac{1}{2}\log(2) + \frac{1}{2}\log(2) = \frac{3}{2}\log(2)$$

In general the entropy of X is an upper bound on the average number of questions needed to find out X, with the unit: 1 question = 1 bit of information = $\log(2)$.

1.2 Mutual information between two discrete random variables

The brain is typically viewed as gathering information about the outside world through its sensors. One way to model this is to consider that the events s (for stimulus) in the outside world are stochastic and that the activity r (for response) of the neurons is correlated with the exterior events. The mutual information between stimulus and response quantifies how much observing one tells you about the other:

$$I(s,r) = H(s) - H(s|r) = H(r) - H(r|s) = H(r) + H(s) - H(s,r) = \sum_{s,r} p(s,r) \log \left[\frac{p(s,r)}{p(s)p(r)} \right]$$

Answer the following questions without calculation (give the answer in bits):

1. What is the mutual information between stimulus and response if they are uncorrelated?

   r and s are uncorrelated $\Leftrightarrow p(s,r) = p(s)p(r) \quad \forall s, r$

   $\Leftrightarrow \log \left[\frac{p(s,r)}{p(s)p(r)} \right] = \log(1) = 0 \quad \forall s, r$

2. Consider a binary stimulus ($s = A, B$), and a single binary neuron with activity $r = 0$ if $s = A$ and $r = 1$ if $s = B$. If both stimuli have the same probability, what is the mutual information between s and the neural activity?

   For a given activity $r = 0, 1$, then s is entirely known. Therefore, the entropy of s given r is null (no information left to know, using the previous analogy there is 0 questions to ask): $H(s|r) = 0$. Both stimuli have same probability, then $H(s) = \log(2) \Rightarrow I(s,r) = \log(2)$.

3. Consider now two neurons, with activities $r_i(s)$, $i = 1, 2$. Suppose that, if $s = A$, $r_1(A) = r_2(A) = 1$, and if $s = B$, $r_1(B) = r_2(B) = 0$. Obviously the neural code is redundant. We want to quantify this redundancy.

   What is the mutual information $I(s, \{r_1, r_2\})$? Between the stimulus and only neuron 1, $I(s, r_1)$?

   One way to define the redundancy is $R \equiv I(s, r_1) + I(s, r_2) - I(s, \{r_1, r_2\})$. What is the redundancy in this particular case?

   As before, if r_1 and r_2 are known then s is entirely known. Even further, if one of r_1, r_2 is known then the other is known as well. We thus have $H(s|r_1, r_2) = 0 = H(s|r_1) = H(s|r_2)$. Again, $H(s) = \log(2)$ such that $I(s|r_1, r_2) = I(s|r_1) = I(s|r_2) = \log(2)$. The redundancy is $R = \log(2)$, which is equal to the mutual information, meaning the whole information is redundant between the two neurons, as expected.

1.3 Mutual information for continuous random variables

Entropy and mutual information can be extended to the case of continuous random variables (in which the sum becomes an integral), according to:

$$H(s) = \int \mathrm{d}s \, p(s) \log [p(s)] \tag{1}$$

$$I(s,r) = \int \mathrm{d}s \, \mathrm{d}r \, p(s,r) \log \left[\frac{p(s,r)}{p(s)p(r)} \right] \tag{2}$$

1. Suppose the neural response is then processed: $s \to r \to x$. Show that processing cannot increase information: $I(s,x) \leq I(s,r)$ (hint: use the concavity of the logarithm). **TODO**

2. Compute the entropy of the Gaussian distribution:

$$P(r) = \frac{1}{(2\pi\sigma^2)^{1/2}}\, e^{-\frac{(r-r_0)^2}{2\sigma^2}}$$

$$\begin{aligned} H(r) = -\int \mathrm{d}r P(r)\log\left[P(r)\right] &= \int \mathrm{d}r P(r)\left[\frac{(r-r_0)^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)\right] \\ &= \frac{1}{2}\left[\langle\frac{(r-r_0)^2}{\sigma^2}\rangle + \log(2\pi\sigma^2)\right] \\ &= \frac{1}{2}\left[1+\log(2\pi\sigma^2)\right] \end{aligned}$$

3. Consider the simple linear system,

$$r = WS + z$$

where S is a scalar stimulus, W is a (positive) weight, z is a Gaussian noise with zero mean and variance σ^2:

$$P(z) = \frac{1}{(2\pi\sigma^2)^{1/2}}\, e^{-\frac{z^2}{2\sigma^2}}$$

and the stimulus has also a Gaussian distribution,

$$\rho(S) = \frac{1}{(2\pi c^2)^{1/2}}\, e^{-\frac{S^2}{2c^2}}$$

Compute the mutual information $I(r,S)$ between the stimulus S and the neural response r.

We can calculate the mutual information in three equivalent ways:

- *First way directly calculating $I(s,r) = H(r) - H(r|s)$.*
  $p(r|s)$ is distributed as a gaussian of mean Ws and variance σ^2, therefore:

$$H(r|s) = \frac{1}{2}\left[1+\log(2\pi\sigma^2)\right]$$

  $p(r)$ is distributed as a gaussian of mean 0 and variance $W^2c^2+\sigma^2$, therefore:

$$H(r) = \frac{1}{2}\left[1+\log(2\pi(W^2c^2+\sigma^2))\right]$$

  Then:

$$\begin{aligned} I(s,r) &= \frac{1}{2}\left[-1-\log(2\pi\sigma^2)+1+\log(2\pi(W^2c^2+\sigma^2))\right] \\ &= \frac{1}{2}\log\left(1+\frac{W^2c^2}{\sigma^2}\right) \end{aligned}$$

- *Second way using $I(s,r) = H(s) - H(s|r)$.*

$$\begin{aligned} H(s) &= \frac{1}{2}\left[1+\log(2\pi c^2)\right] \\ p(s|r) = \frac{p(r|s)p(s)}{p(r)} &\propto p(r|s)p(s) \\ &\propto e^{-\frac{s^2}{2c^2}-\frac{(r-Ws)^2}{2\sigma^2}} \\ &\propto e^{-s^2(\frac{1}{2c^2}+\frac{W^2}{2\sigma^2})} \\ &\propto e^{-\frac{s^2}{2\frac{1}{\frac{1}{c^2}+\frac{W^2}{\sigma^2}}}} \end{aligned}$$

Indeed $p(s|r)$ is a product of Gaussians, therefore the inverse variances add.

$$\begin{aligned} H(s|r) &= \frac{1}{2}\left[1+\log\left(2\pi\frac{1}{\frac{1}{c^2}+\frac{W^2}{\sigma^2})}\right)\right] \\ I(s,r) &= \frac{1}{2}\left[1+\log(2\pi)+\log(c^2)-1-\log(2\pi)+\log\left(\frac{1}{c^2}+\frac{W^2}{\sigma^2}\right)\right] \\ &= \frac{1}{2}\log\left(1+\frac{c^2W^2}{\sigma^2}\right) \end{aligned}$$

- *Third way using $I(s,r)=H(s)+H(r)-H(s,r)$.*
  For a multivariate Gaussian of covariance matrix Σ:

$$p(\vec{X}) = \frac{1}{\sqrt{(2\pi)^N|\Sigma|}}\exp(-(\vec{X}-\vec{X}_0)^T\Sigma^{-1}(\vec{X}-\vec{X}_0))$$

The entropy is given by:

$$H(\vec{X}) = \log\left(\sqrt{(2\pi e)^N|\Sigma|}\right)$$

Here $\vec{X}=(r,s)$, the covariance matrix is given by:

$$\begin{aligned} r^2 = W^2s^2+z^2+Wsz \quad &\Rightarrow \quad \langle r^2\rangle = W^2c^2+\sigma^2 \\ rs = Ws^2+sz \quad &\Rightarrow \quad \langle rs\rangle = Wc^2 \\ \Sigma = \begin{pmatrix} \langle r^2\rangle & \langle rs\rangle \\ \langle rs\rangle & \langle s^2\rangle \end{pmatrix} \quad &\Rightarrow \quad \Sigma = \begin{pmatrix} W^2c^2+\sigma^2 & Wc^2 \\ Wc^2 & c^2 \end{pmatrix} \end{aligned}$$

Its determinant is $|\Sigma| = c^2\sigma^2$. Then:

$$\begin{aligned} H(r,s) = 1+\log(2\pi\sigma) \quad \Rightarrow \quad I(s,r) &= \frac{1}{2}\left[1+\log(2\pi c^2)\right]+\frac{1}{2}\left[1+\log(2\pi(W^2c^2+\sigma^2))\right]-1-\log(2\pi c\sigma) \\ &= \frac{1}{2}\log\left[\frac{c^2(W^2c^2+\sigma^2)}{c^2\sigma^2}\right] \\ &= \frac{1}{2}\log\left[1+\frac{W^2c^2}{\sigma^2}\right] \end{aligned}$$

4. Compute the mutual information $I(r,S)$ for the model

$$r = \sum_{j=1}^{N} W_jS_j + z$$

with Gaussian inputs of zero mean, $\langle S_j\rangle = 0$ for every j, and covariance matrix C:

$$\langle S_jS_{j'}\rangle = C_{j,j'}$$

Discuss the maximization of mutual information with respect to the choice of the weights $\{W_j, j = 1,...,N\}$.

r is a sum of Gaussians and is therefore also a Gaussian.

$$
\begin{aligned}
\langle r \rangle &= \sum_{j=1}^{N} w_j \langle s_j \rangle + \langle z \rangle = 0 \\
\langle r^2 \rangle &= \langle \sum_{i,j=1}^{N} w_i w_j s_i s_j + 2 \sum_{j=1}^{N} w_j s_j z + z^2 \rangle \\
&= \sum_{i,j=1}^{N} w_i w_j c_{i,j} + \sigma^2 = \vec{W}^T C \vec{W} + \sigma^2 \\
H(r) &= \frac{1}{2} \left[1 + \log(2\pi(\vec{W}^T C \vec{W} + \sigma^2)) \right] \\
H(r|s) &= \frac{1}{2} \left[1 + \log(2\pi\sigma^2) \right] \\
I(r,s) &= H(r) - H(r|s) = \frac{1}{2} \log \left[1 + \frac{\vec{W}^T C \vec{W}}{\sigma^2} \right]
\end{aligned}
$$

Since C is symmetrical and real, it can be diagonalized in an orthonormal basis. We denote $w_{j,new}$ the coordinates of $\vec{W}$ in this new basis and λ_j the eigenvalues of C.

$$
\begin{aligned}
\vec{W}^T C \vec{W} &= \sum_j \lambda_j w_{j,new}^2 \\
\|\vec{W}\|^2 &= \sum_j w_{j,new}^2 = 1
\end{aligned}
$$

The mutual information is therefore maximal when $\vec{W}$ is aligned with the eigenvector of C associated with the maximum eigenvalue $\lambda_{j,max}$. Indeed, since the noise has the same variance in all directions, the direction with the best signal to noise ratio is the one where the signal has the highest variance.

2 Fisher Information

Shannon's information is a measure of how much information the response of a neuron provides about the whole stimulus space. However, individual neurons in the brain appear to be "tuned" to certain regions of the stimulus space. It can therefore be useful to introduce a local measure of information, such as the Fisher information, which can be characterized in two equivalent ways.

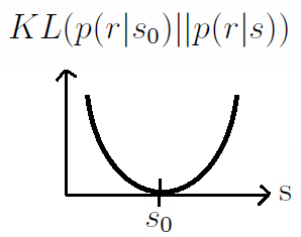
2.1 Distance between probability distributions

The 'precision' of the information that a neuron provides about a stimulus s_0 corresponds to how easy it is, given the response of that neuron, to tell s_0 apart from nearby values $s_0 + \delta s$. The more the distributions $p(r|s_0)$ and $p(r|s_0 + \delta s)$ overlap, the less 'precise' this information is. To quantify the distance between these two distributions, we introduce the Kullback-Leibler divergence:

$$
\begin{aligned}
KL(p||q) &= \int \mathrm{d}x \, p(x) \log \left[\frac{p(x)}{q(x)} \right] && (3) \\
KL(p||q) &\geq 0 && (4) \\
KL(p||q) = 0 &\Leftrightarrow p = q && (5)
\end{aligned}
$$

Note that this distance has already been used to define the mutual information between s and r as the distance between $p(s,r)$ and $p(s)p(r)$.

1. Sketch $KL(p(r|s_0)||p(r|s))$ as a function of s.



2. Explain why the following gives a measure of the information that r locally provides about the stimulus:

$$F(s_0) = \frac{d^2 KL(p(r|s_0)||p(r|s))}{ds^2}(s_0) = -\int \mathrm{d}r\, p(r|s_0) \frac{\partial}{\partial S} \log(p(r|s))(s_0)$$

The second derivative at s_0 provides a measure of "how fast" the distributions separate out.

2.2 Variance of the locally optimal estimator

Another approach to quantify the 'precision' of the information that the neural response provides about the stimulus is to try to estimate the stimulus from the response, using an estimator $\hat{s}(r)$. We only consider locally unbiased estimators; estimators which are accurate on average for values of the stimulus close to s_0:

$$\frac{\partial}{\partial S}\langle \hat{s} \rangle (s_0) = 1$$

where $\langle \hat{s} \rangle (s) = \int \hat{s}(r) p(r|s)\, \mathrm{d}r$.

1. Such an estimator, although accurate on average, will generally not provide an exact estimate of the stimulus on each trial. How could we measure the 'precision' of such an estimator?

   Using its variance at s_0:

$$\int \mathrm{d}r\, p(r|s_0)(\hat{s}(r) - s_0)^2$$

2. Using the fact that the estimator is unbiased, and the Cauchy-Schwarz inequality:

$$\int f(x)g(x)\,\mathrm{d}x \leq \int f^2(x)\,\mathrm{d}x \int g^2(x)\,\mathrm{d}x$$

   with equality if and only if $f(x) = ag(x)$, show that:

$$\int \mathrm{d}r\, p(r|s_0)(\hat{s}(r) - s_0)^2 \int \mathrm{d}r\, p(r|s_0) \left(\frac{\partial}{\partial S} \log(p(r|s))(s_0) \right)^2 \geq 1$$

   We define $f(r) = \sqrt{p(r|s_0)}(\hat{s}(r) - s_0)$ and $g(r) = \sqrt{p(r|s_0)} \frac{\partial}{\partial S} log(p(r|s))(s_0)$:

$$\begin{aligned}
\int f(r)g(r)\,\mathrm{d}r &= \int \mathrm{d}r\, p(r|s_0)(\hat{s}(r) - s_0) \frac{\partial}{\partial S} log(p(r|s))(s_0) \\
&= \int \mathrm{d}r (\hat{s}(r) - s_0) \frac{\partial}{\partial S} p(r|s)(s_0) \\
&= \frac{\partial}{\partial S} \left[\int dr (\hat{s}(r) - s_0) p(r|s))(s_0) \right] \\
&= \frac{\partial}{\partial S} (\langle \hat{s}(r) - s_0 \rangle)(s_0) = 1
\end{aligned}$$

The last equality comes from having an unbiased estimator. We apply the Cauchy-Schwarz inequality:

$$1 \leq \int \mathrm{d}r\, p(r|s_0)(\hat{s}(r) - s_0)^2 \int \mathrm{d}r\, p(r|s_0) \left[\frac{\partial}{\partial S} \log(p(r|s))(s_0)\right]^2$$

Using the fact that the probability distribution is normalized, it is possible to show that this limit which we obtain on the 'precision' of the information is equivalent to the Fisher Information defined in the previous section:

$$F(s) = \int \mathrm{d}r\, p(r|s) \left[\frac{\partial}{\partial S} \log p(r|s)\right]^2 = -\int \mathrm{d}r\, p(r|s) \frac{\partial^2}{\partial S^2} \log p(r|s)$$

3. Using the case of equality in the Cauchy-Schwarz inequality, find a locally unbiased estimator whose variance is equal to the inverse of the Fisher Information.

   For $f(r) = ag(r)$:

   $$\int f(r)g(r)\,\mathrm{d}r = \int f^2(r)\,\mathrm{d}r \int g^2(r)\,\mathrm{d}r$$

   We therefore consider $\hat{s}(r) - s_0 = a\frac{\partial}{\partial S} log(p(r|s))(s_0)$.

   Note that the estimator doesn't depend on s, its mean value depends on s only through $p(r|s)$. Therefore, we use the fact that the estimator is unbiased:

   $$\begin{aligned} 1 &= \frac{\partial}{\partial S}\Big(\int (\hat{s}(r) - s_0)p(r|s)\,\mathrm{d}r\Big)(s_0) \\ &= a\Big(\int \mathrm{d}r\, \frac{\partial}{\partial S} p(r|s) \frac{\partial}{\partial S} \log(p(r|s))(s_0)\Big)(s_0) \\ &= a\int \mathrm{d}r\, \frac{\partial}{\partial S} p(r|s)(s_0) \frac{\partial}{\partial S} \log(p(r|s))(s_0) \end{aligned}$$

   With $\frac{\partial}{\partial S} \log(p(r|s)) = \frac{\frac{\partial}{\partial S} p(r|s)}{p(r|s)}$ we therefore have:

   $$1 = a\int \mathrm{d}r\, p(r|s_0)\Big(\frac{\partial}{\partial S} \log(p(r|s))(s_0)\Big)^2 = aF(s_0) \quad \Rightarrow \quad a = 1/F(s_0)$$

   We can verify that this is the right constant by checking that the variance of the estimator that we find is indeed equal to the inverse of the Fisher information:

   $$\langle(\hat{s}(r) - s_0)^2\rangle = a^2 F(s_0) = 1/F(s_0)$$

2.3 Some examples

1. Suppose that the mean response of a neuron to a stimulus s is $f(s)$ and the variance of the neuron's response is $\sigma(s)^2$. How do you expect the Fisher Information to depend on $f(s)$ and $\sigma(s)$?

   The Fisher information is the inverse of the variance of an estimator of s, its unit is therefore $1/[s]^2$. The unit of σ is $[f]$ and the unit of $f'(s)$ is $[f]/[s]$, therefore by dimensionality analysis:

   $$F(s) \propto \Big(\frac{f'(s)}{\sigma(s)}\Big)^2$$

   Indeed the Fisher Information is comparable to the signal to noise ratio.

2. Give the Fisher Information for a neuron with a Poisson firing rate:

$$P(r|s) = \frac{f(s)^r}{r!}\, e^{-f(s)}$$

What is the optimal estimator?

From the previous results, we want $\hat{s}(r) - s_0 = \frac{1}{F(s_0)} \frac{\partial}{\partial S} log(p(r|s))(s_0)$.

We have $\log(p(r|s)) = r\log(\lambda(s)) - \lambda(s) + \log(r!)$ *such that:*

$$\frac{\partial}{\partial S}\log(p(r|s))(s_0) = r\frac{\lambda'(s_0)}{\lambda(s_0)} - \lambda'(s_0) = \frac{\lambda'(s_0)}{\lambda(s_0)}(r - \lambda(s_0))$$

The Fisher information is:

$$\begin{aligned} F(s_0) &= \int \mathrm{d}r\, p(r|s_0)[\frac{\partial}{\partial S}\log p(r|s)]^2 \\ &= (\frac{\lambda'(s_0)}{\lambda(s_0)})^2 \int \mathrm{d}r\, p(r|s_0)(r - \lambda(s_0))^2 \\ &= \frac{\lambda'(s_0)^2}{\lambda(s_0)} \end{aligned}$$

The optimal estimator follows:

$$\hat{s}(r) - s_0 = \frac{\lambda(s_0)}{\lambda'(s_0)^2}\frac{\lambda'(s_0)}{\lambda(s_0)}(r - \lambda(s_0)) = \frac{r - \lambda(s_0)}{\lambda'(s_0)}$$

3. What is the Fisher Information for a neuron with Gaussian noise:

$$P(r|s) = \frac{1}{(2\pi)^{1/2}\sigma(s)}\exp(-\frac{(r - f(s))^2}{2\sigma(s)^2})$$

The same developments give:

$$\begin{aligned} &\log(p(r|s)) = -\frac{(r - f(s))^2}{2\sigma(s)^2} - \log((2\pi)^{1/2}) - \log(\sigma(s)) \\ \Rightarrow\quad &\frac{\partial}{\partial S}\log(p(r|s))(s_0) = \frac{f'(s_0)(r - f(s_0))}{\sigma(s_0)^2} + (r - f(s_0))^2\frac{4\sigma'(s_0)\sigma(s_0)}{4\sigma(s_0)^4} - \frac{\sigma'(s_0)}{\sigma(s_0)} \end{aligned}$$

with $\int \mathrm{d}r\, p(r|s_0) = 1$, $\int \mathrm{d}r\, p(r|s_0)(r - f(s_0)) = 0$ *and* $\int \mathrm{d}r\, p(r|s_0)(r - f(s_0))^2 = \sigma(s_0)^2$, *such that* $\int \mathrm{d}r\, p(r|s_0)(r - f(s_0))^3 = \int \mathrm{d}r\, p(r|s_0)(r - f(s_0))^4 = 0$. *Then:*

$$F(s) = \frac{f'(s_0)^2}{\sigma(s_0)^2} + 0 + \left(\frac{\sigma'(s_0)}{\sigma(s_0)}\right)^2 + 0 - \sigma(s_0)^2\frac{\sigma'(s_0)}{\sigma(s_0)^3}\frac{\sigma'(s_0)}{\sigma(s_0)} - 0 = \frac{f'(s_0)^2}{\sigma(s_0)^2}$$

The optimal estimator is:

$$\begin{aligned} \hat{s}(r) - s_0 &= \frac{\sigma(s_0)^2}{f'(s_0)^2}\left[\frac{f'(s_0)(r - f(s_0))}{\sigma(s_0)^2} + (r - f(s_0))^2\frac{4\sigma'(s_0)\sigma(s_0)}{4\sigma(s_0)^4} - \frac{\sigma'(s_0)}{\sigma(s_0)}\right] \\ &= \frac{r - f(s_0)}{f'(s_0)} + \frac{\sigma'(s_0)}{\sigma(s_0)}\frac{(r - f(s_0))^2 - \sigma(s_0)^2}{f'(s_0)^2} \end{aligned}$$

Supposing that the variance is constant, $\sigma'(s_0) = 0$, *then again the optimal estimator is given by:*

$$\hat{s}(r) - s_0 = \frac{r - f(s_0)}{f'(s_0)}$$

4. What is the Fisher Information for two independent neurons?

   What is the optimal estimator? For simplicity, consider that the variance is constant: $\sigma'(s) = 0$

   The Fisher Information for two independent neurons is the sum of the Fisher information of each neuron because:

$$p(r_1, r_2|s) = p(r_1|s)p(r_2|s)$$
$$\log(p(r_1, r_2|s)) = \log(p(r_1|s)) + \log(p(r_2|s))$$