TD 10 – Neuronal Coding & Information Theory

1 Mutual Information

1.1 Characterizing the distribution of a discrete random variable

① *Entropy & Average number of questions*

• Only one possible color.

Number of questions : It is not necessary to ask any question since the color of the ball is certain.
Entropy : $H = -1\log(1) = 0$.

• Two possible colors.

Number of questions : One question is sufficient, for instance "Is it red ?" if the possible colors are blue and red.
Entropy : $H = -p\log(p) - (1-p)\log(1-p)$.
- $H = 0$ for $p = 0$ or 1
- $H = \log(2)$ is maximal at $p = 0.5$.

• Half the balls are red, one fourth are green and one fourth are blue.

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

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

• Comment : In general the entropy of a random variable X is an upper bound on the average number of questions needed to find out the value of X in a given event. The 'unit' of the entropy is in *bits of information* : 1 question $= 1$ bit of information $= \log(2)$.

1.2 Mutual information between two discrete random variables

② *Mutual information between uncorrelated stimulus and response*

r and s are uncorrelated implies that $p(s,r) = p(s)p(r) \quad \forall s, r$.
Thus : $\log\left(\frac{p(s,r)}{p(s)p(r)}\right) = \log(1) = 0 \quad \forall s, r$

③ *Mutual information for binary stimuli with identical probability*

For a given activity $r = 0, 1$, then the stimulus 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 $H(s) = \log(2)$.
Thus : $I(s,r) = \log(2)$.

④ *Mutual information for two neurons*

• For two neurons : 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. Thus : $H(s|r_1, r_2) = 0 = H(s|r_1) = H(s|r_2)$.
Again, $H(s) = \log(2)$. Thus : $I(s|r_1, r_2) = I(s|r_1) = I(s|r_2) = \log(2)$. • For one neuron :

⑤ *Redundancy*

$R = \log(2)$, which is equal to the mutual information. This means that the whole information is redundant between the two neurons, as expected.

1.3 Mutual information for continuous random variables

⑥ *Entropy of the Gaussian distribution*

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

⑦ *Mutual information for a Gaussian stimulus with Gaussian noise*

- Method 1 : Using $I(s,r) = H(r) - H(r|s)$.

$p(r|s)$ 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 :

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

Thus :

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

- Method 2 : Using $I(s,r) = H(s) - H(s|r)$.

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

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

$$p(s|r) = \frac{p(r|s)p(s)}{p(r)} \propto p(r|s)p(s)$$

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

Thus :

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

- Method 3 : 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 :

$$H(\vec{X}) = \log\left(1 + \sqrt{(2\pi e)^N|\Sigma|}\right)$$

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

$$r^2 = w^2s^2 + z^2 + wsz \quad\Rightarrow\quad \langle r^2\rangle = w^2c^2 + \sigma^2$$

$$rs = ws^2 + sz \quad\Rightarrow\quad \langle rs\rangle = wc^2$$

$$\Sigma = \begin{pmatrix} \langle r^2\rangle & \langle rs\rangle \\ \langle rs\rangle & \langle s^2\rangle \end{pmatrix} \quad\Rightarrow\quad \Sigma = \begin{pmatrix} w^2c^2 + \sigma^2 & wc^2 \\ wc^2 & c^2 \end{pmatrix}$$

Its determinant is $|\Sigma| = c^2\sigma^2$. (Note : The formula for the variances stem from the null mean).

Thus :

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

8 *Mutual information for the model of Gaussian inputs with covariance matrix*

r is Gaussian, as a sum of Gaussians.

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

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

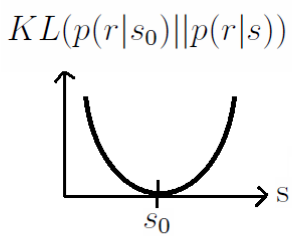
$$\vec{W}^T C\vec{W} = \sum_j \lambda_j w_{j,new}^2$$
$$\|\vec{W}\|^2 = \sum_j w_{j,new}^2 = 1$$

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

2.1 Distance between probability distributions

9 *Sketch of the Kullbach-Liebler divergence*



10 *Measure of local information*

The second derivative at s_0 provides a measure of "how fast" both distributions separate out. The second derivative is used instead of the first derivative because the latter cancels at s_0 :

$$\begin{aligned} F(s_0) &= \left.\frac{\partial^2 KL(p(r|s_0)||p(r|s))}{\partial s^2}\right|_{s_0} \\ &= \left.\left[\frac{\partial^2}{\partial s^2}\int p(r|s_0)\log\left(\frac{p(r|s_0)}{p(r|s)}\right)dr\right]\right|_{s_0} \\ &= \left.\left[\frac{\partial^2}{\partial s^2}\int p(r|s_0)\log(p(r|s_0))dr\right]\right|_{s_0} - \left.\left[\frac{\partial^2}{\partial s^2}\int p(r|s_0)\log(p(r|s))dr\right]\right|_{s_0} \end{aligned}$$

The first term does not depend on s and thus cancels, such that : $F(s_0) = -\int p(r|s_0)\ \left.\left[\frac{\partial^2}{\partial s^2}\log(p(r|s))\right]\right|_{s_0} dr$

2.2 Variance of the locally optimal estimator

11 *Unbiased estimator*

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

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

The second last equality stems from :

$\left.\frac{\partial\log(p(r|s)}{\partial s}\right|_{s_0} = \left.\frac{\frac{\partial p(r|s)}{\partial s}}{p(r|s)}\right|_{s_0} = \frac{\left.\frac{\partial p(r|s)}{\partial s}\right|_{s_0}}{p(r|s_0)}$ such that $p(r|s_0)\ \left.\frac{\partial\log(p(r|s)}{\partial s}\right|_{s_0} = \left.\frac{\partial p(r|s)}{\partial s}\right|_{s_0}$

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

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

12 *Equality between two formulas*

On the one hand : $\int p(r|s_0)\left(\left.\frac{\partial}{\partial s}\log(p(r|s))\right|_{s_0}\right)^2 dr = \int p(r|s_0)\left(\left.\frac{\frac{\partial}{\partial s}p(r|s)}{p(r|s)}\right|_{s_0}\right)^2 dr$

$$\begin{aligned} &= \int p(r|s_0)\left(\frac{\frac{\partial}{\partial s}p(r|s)\big|_{s_0}}{p(r|s_0)}\right)^2 dr \\ &= \int \frac{\left(\frac{\partial}{\partial s}p(r|s)\big|_{s_0}\right)^2}{p(r|s_0)}dr \end{aligned}$$

On the other hand : $-\int p(r|s_0)\ \left.\frac{\partial^2}{\partial s^2}\log(p(r|s))\right|_{s_0} dr = -\int p(r|s_0)\ \left.\frac{\partial}{\partial s}\left(\frac{\frac{\partial}{\partial s}p(r|s)}{p(r|s)}\right)\right|_{s_0} dr$

$$\begin{aligned} &= -\int p(r|s_0)\ \left.\frac{\frac{\partial^2}{\partial s^2}p(r|s)\times p(r|s) - \left(\frac{\partial}{\partial s}p(r|s)\right)^2}{p(r|s)^2}\right|_{s_0} dr \\ &= -\int p(r|s_0)\frac{\frac{\partial^2}{\partial s^2}p(r|s)\big|_{s_0}\ p(r|s_0) - \left(\frac{\partial}{\partial s}p(r|s)\big|_{s_0}\right)^2}{p(r|s_0)^2}dr \\ &= -\underbrace{\int \left.\frac{\partial^2}{\partial s^2}p(r|s)\right|_{s_0} dr}_{0} + \int \frac{\left(\frac{\partial}{\partial s}p(r|s)\big|_{s_0}\right)^2}{p(r|s_0)}dr \end{aligned}$$

The first term vanishes because : $\frac{\partial^2}{\partial s^2}\int p(r|s)dr\Big|_{s_0} = \frac{\partial^2}{\partial s^2}1\Big|_{s_0} = 0$.

(13) *Locally unbiased estimator whose variance is equal to the inverse of the Fisher Information*

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

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

Let us therefore consider $\widehat{s}(r) - s_0 = a\frac{\partial}{\partial S}log(p(r|s))(s_0)$. Note that the estimator does not depend on s, its mean value depends on s only through $p(r|s)$.
Using the fact that the estimator is unbiased :

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

Using $\frac{\partial}{\partial S}\log(p(r|s)) = \frac{\frac{\partial}{\partial S}p(r|s)}{p(r|s)}$:

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

It can be verified that this is the right constant by checking that the variance of the obtained estimator is indeed equal to the inverse of the Fisher information :

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

2.3 Examples of Fisher local information for different response models

(14) *Dependence of Fisher Information on mean and variance*

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 :

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

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

(15) *Neuron with Poisson firing rate*

From the previous results, the optimal estimator is given by : $\widehat{s}(r) - s_0 = \dfrac{1}{F(s_0)}\dfrac{\partial}{\partial S}\log(p(r|s))(s_0)$.
In this example : $\log(p(r|s)) = r\log(\lambda(s)) - \lambda(s) + \log(r!)$ 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 :

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

The optimal estimator follows :

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

(16) *Neuron with Gaussian noise*

The same developments give :

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

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 :

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

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

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

(17) *Two independent neurons*

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

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

3 Bayesian inference

(18) *Probability of observing a pattern of spikes*

$$\mathbb{P}(\{n_i\}|s) = \prod_{i=1}^{N}\frac{f_i(s)^{n_i}}{n_i!}\exp(-f_i(s))$$

(19) *Estimate of the stimulus*

An estimate of the stimulus can be built by using the intuition that each neuron 'votes' for its preferred stimulus, such that the estimate is a weighted average :

$$\frac{\sum_{i=1}^N n_i s_i}{\sum_{i=1}^N n_i}$$

Note that the same pattern can be caused by various stimuli, therefore by observing the pattern cannot allow to infer unambiguously the exact stimulus.

(20) *Effect of the tuning curves on accuracy*

Using previous results, the Fisher information is

$$\sum_{i=1}^{N}\frac{f_i'(s)^2}{f_i(s)} \approx N\frac{(f_0/\sigma)^2}{f_0} = \frac{Nf_0}{\sigma^2}$$

. This formula confirms the intuitions that the local accuracy improves with more neurons and tuning curves with higher-magnitudes, whereas is it impairs with the variance of tuning curves (which contribute to greater overlap and more confusion).

21 *Probability distribution of the stimulus using Baye's rule*

$$p(s|\{n_i\}) = \frac{p_s}{p(\{n_i\})} \prod_{i=1}^{N} \frac{1}{n_i!} \exp\left(-\sum_{i=1}^{N} f_i(s)\right) \exp\left(\sum_{i=1}^{N} n_i \log(f_i(s))\right) \text{ where } p_s = p(s) \forall s$$

The tuning curves being evenly distributed along the stimulus range, it can be considered that $\sum_{i=1}^{N} f_i(s)$ does not depend on s.

$$p(s|\{n_i\}) = \Phi(\{n_i\}) \exp\left(\sum_{i=1}^{N} n_i \log(f_i(s))\right) \tag{1}$$

22 *Gaussian tuning curves*

The estimate of the stimulus corresponds to the maximum of $\log(p(\{n_i\}|s))$:

$$\frac{\partial}{\partial s} \log(p(\{n_i\}|s)) = \sum_{i=1}^{N} n_i \left(-2\frac{(s-s_i)}{2\sigma^2}\right) = 0 \Leftrightarrow \sum_{i=1}^{N} n_i s = \sum_{i=1}^{N} n_i s_i \Leftrightarrow s = \frac{\sum_{i=1}^{N} n_i s_i}{\sum_{i=1}^{N} n_i}$$

The distribution of the stimulus is given by :

$$p(s|\{n_i\}) = \Phi(\{n_i\}) \exp\left(\sum_{i=1}^{N} -n_i \frac{(s-s_i)^2}{2\sigma^2}\right)$$

The quantity inside the exponential rewrites :

$$\sum_{i=1}^{N} n_i (s-s_i)^2 = \left(\sum_{i=1}^{N} n_i\right) s^2 - 2\left(\sum_{i=1}^{N} n_i s_i\right) s + \sum_{i=1}^{N} n_i s_i^2 = \left(\sum_{i=1}^{N} n_i\right)\left(s - \frac{\sum_{i=1}^{N} n_i s_i}{\sum_{i=1}^{N} n_i}\right)^2 + C$$

This implies that the distribution is Gaussian with the following parameters :

$$\text{Mean : } \frac{\sum_{i=1}^{N} n_i s_i}{\sum_{i=1}^{N} n_i} \qquad \text{Variance : } \frac{\sigma^2}{\sum_{i=1}^{N} n_i}$$

The variance of the posterior depends on the sum of the n_i, which is proportional to the number of neurons and the height of the tuning curves, therefore the variance is proportional to $\frac{\sigma^2}{N f_0}$.
The variance becomes infinitely small if the number of neurons becomes infinitely large or if the mean firing rate becomes infinitely large.

23 *Jitterred responses*

The variability is now correlated across neurons :

$$p(s|\{n_i\}) = p(s|\widehat{s}) p(\widehat{s}|\{n_i\})$$

It is now a product of Gaussians : the inverse variances add. In the previous conditions, $p(\widehat{s}|\{n_i\})$ becomes infinitely narrow, its variance becomes infinitely small. However the variance of $p(s|\{n_i\})$ is necessarily larger than the variance of $p(s|\widehat{s})$ which does not depend on the number of neurons or on their mean firing rate.