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 color.

Number of questions : It is not necessary to ask any question since the color of the ball is certain.
Entropy :

$$H = -1\log(1) = 0$$

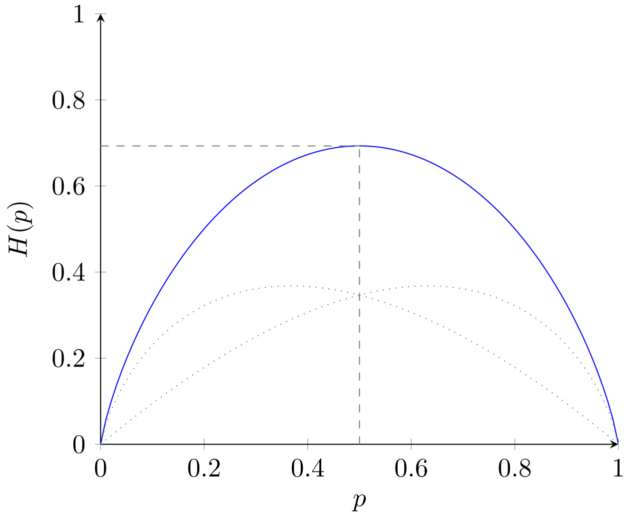
- Two colors.

Number of questions : One question is sufficient. For instance, if the possible colors are blue and red : "Is it 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 : At least one question ("Is it red ?") and at most two questions ("Is it red ?", "Is it blue ?") suffice to determine the color.
One question is sufficient in half of cases (i.e. when the ball is red), and two questions are necessary in the other cases. Therefore, in average, the number of question is equal to : $\frac{1}{2} \times 1 + \frac{1}{2} \times 2 = \frac{3}{2}$.

Entropy :

$$\begin{aligned} H &= -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{4}\log\left(\frac{1}{4}\right) - \frac{1}{4}\log\left(\frac{1}{4}\right) \\ &= \frac{1}{2}\log(2) + 2 \times \frac{1}{4}\log(4) \\ &= \frac{1}{2}\log(2) + 2 \times \frac{1}{4}\log(2^2) \\ &= \frac{1}{2}\log(2) + 2 \times \frac{1}{4} \times 2\log(2) \\ &= \frac{1}{2}\log(2) + \log(2) \\ &= \frac{3}{2}\log(2) \end{aligned}$$

② Comment :

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

1.2 Mutual information between two discrete random variables

③ *Mutual information between uncorrelated stimulus and response*

The assumption that r and s are uncorrelated entails that $p(s,r) = p(s)p(r) \quad \forall s, r$.

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

④ *Mutual information for binary stimuli with identical probability*

Both stimuli have same probability, then $H(s) = \log(2)$.
For any activity $r \in \{0,1\}$, then the stimulus s is entirely known : using the previous analogy, not any question is necessary. Therefore, the entropy of s given r is null : $H(s|r) = 0$.

$$I(s,r) = H(s) - H(s|r) = \log(2)$$

⑤ *Mutual information for two neurons*

- With the joint responses of both neurons :

As previously, $H(s) = \log(2)$.
As previously, for any joint activity $\{r_1, r_2\}$, then the stimulus s is entirely known : $H(s|r_1, r_2) = 0$

$$I(s|r_1, r_2) = \log(2)$$

- With the response of one neuron :

As soon as one of r_1, r_2 is known, then the other is known as well, and so does the stimulus : $H(s|r_1) = H(s|r_2) = 0$.

$$I(s|r_1) = I(s|r_2) = \log(2)$$

⑥ *Redundancy*

$$I(s,r_1) + I(s,r_2) - I(s,\{r_1,r_2\}) = 2 \times \log(2) - \log(2) = \log(2)$$

In this case, the redundancy 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

⑦ *Entropy of the Gaussian distribution*

$$\begin{aligned}
H(r) &= -\int \mathrm{d}r P(r)\log\left[P(r)\right] = -\int \mathrm{d}r P(r)\log\left(\frac{1}{(2\pi\sigma^2)^{1/2}}\, e^{-\frac{(r-r_0)^2}{2\sigma^2}}\right) \\
&= -\int \mathrm{d}r P(r)\left[\log\left(\frac{1}{(2\pi\sigma^2)^{1/2}}\right) - \log\left(e^{-\frac{(r-r_0)^2}{2\sigma^2}}\right)\right] \\
&= \int \mathrm{d}r P(r)\frac{1}{2}\log(2\pi\sigma^2) + \int \mathrm{d}r P(r)\frac{(r-r_0)^2}{2\sigma^2} \\
&= \frac{1}{2}\log(2\pi\sigma^2)\underbrace{\int \mathrm{d}r P(r)}_{1} + \frac{1}{2\sigma^2}\underbrace{\int \mathrm{d}r P(r)(r-r_0)^2}_{\mathbb{V}(r)} \\
&= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sigma^2 \\
&= \frac{1}{2}\left[1 + \log(2\pi\sigma^2)\right]
\end{aligned}$$

⑧ *Mutual information for a Gaussian stimulus with Gaussian noise*

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

Both terms can be computed straightforward by applying the formula of the entropy of a gaussian distribution ⑦. It only requires to specify the variance of the distributions of interest (parameter σ in the formula).

- The random variable r is a sum of two Gaussian variables : ws with variance w^2c^2, and z with variance σ^2. Because s and z are independent, their variances sum, such that the variance of r is $w^2c^2 + \sigma^2$.

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

- The random variable $r|s$ (i.e. for a fixed value of s) has a variance which is only due to z (because s is determined), and thus its variance is equal to σ^2.

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

$\boxed{\text{Conclusion}}$

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

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

- The random variable s has a variance c^2.

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

- ⚠ For the random variable $s|r$, it is not possible to proceed as $r|s$, using the relation $s = \frac{r-z}{w}$ (which would entail that $s|r$ has a variance $\frac{\sigma^2}{w^2}$, as r is fixed). Indeed, the random variables r and z are **not** independent. When the value of r is fixed, it has implications on the possible values of z which could have led to this result.

Therefore, it is necessary to go through the expression of the conditional probability :

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

The computations lead to rewrite this formula under the form of another gaussian distribution, whose variance is an inverse of inverses : $\frac{1}{\frac{1}{c^2} + \frac{w^2}{\sigma^2}}$.

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

$\boxed{\text{Conclusion}}$

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

Method 3 $I(s,r) = H(s) + H(r) - H(s,r)$.

To compute the missing term $H(s,r)$, it is necessary to obtain the probability distribution of the joint events $p(r,s)$.

To do so, the couple of gaussian variables (r,s) can be formalized as a gaussian vector $\vec{X} = \begin{bmatrix} r \\ s \end{bmatrix}$.

It is associated with :

- A mean *vector* $\vec{X}_0 = \begin{bmatrix} \mathbb{E}(r) \\ \mathbb{E}(s) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$
- A covariance *matrix* $\Sigma = \begin{bmatrix} \mathbb{V}(r) & \mathbb{C}\text{ov}(r,s) \\ \mathbb{C}\text{ov}(r,s) & \mathbb{V}(s) \end{bmatrix} = \begin{bmatrix} w^2c^2 + \sigma^2 & wc^2 \\ wc^2 & c^2 \end{bmatrix}$.

Indeed : $\mathbb{C}\text{ov}(r,s) = \mathbb{E}(rs) - \cancel{\mathbb{E}(r)\mathbb{E}(s)} = \mathbb{E}((ws+z)s) = w\underbrace{\mathbb{E}(s^2)}_{\mathbb{V}(s)} + \underbrace{\mathbb{E}(sz)}_{\mathbb{C}\text{ov}(s,z)} = wc^2 + 0$

Note : The last formula for $\mathbb{V}(s)$ and $\mathbb{C}\text{ov}(s,z)$ and stem from the null mean and independence.

For a multivariate Gaussian of N variables with covariance matrix Σ, the probability distribution is given by :

$$p(\vec{X}) = \frac{1}{\sqrt{(2\pi)^N|\Sigma|}}\exp(-(\vec{X} - \vec{X}_0)^T\Sigma^{-1}(\vec{X} - \vec{X}_0)) \quad \text{with}|\Sigma|\text{the determinant of the matrix}$$

The entropy is given by :

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

Here, $N = 2$ and $|\Sigma| = (w^2c^2 + \sigma^2)c^2 - w^2c^4 = c^2\sigma^2$, therefore :

$$H(r,s) = \log\left(\sqrt{(2\pi e)^2 c^2\sigma^2}\right) = \log(2\pi e c\sigma) = 1 + \log(2\pi c\sigma)$$

Conclusion

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

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

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

• As before, the variance of $r|s$ can be determined straightforward : because the stimuli are fixed, it reduces to the variance of z.

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

• The random variable r is Gaussian, as a sum of Gaussians, but the random variables s_i are not independent. Therefore, the variance of r takes into account their covariances :

$$\mathbb{E}(r) = \sum_{j=1}^{N} w_j \mathbb{E}(s_j) + \mathbb{E}(z) = 0$$

$$\begin{aligned} \mathbb{V}(r) = \mathbb{E}(r^2) - \underbrace{\mathbb{E}^2(r)}_{0} &= \mathbb{E}\left(\sum_{i,j=1}^{N} w_i w_j s_i s_j + 2\sum_{j=1}^{N} w_j s_j z + z^2\right) \\ &= \sum_{i,j=1}^{N} w_i w_j c_{i,j} + \sigma^2 \\ &= \vec{W}^T C \vec{W} + \sigma^2 \end{aligned}$$

Hence :

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

Conclusion

$$I(r,s) = \frac{1}{2}\left[1 + \log(2\pi\sigma^2)\right] - \frac{1}{2}\left[1 + \log(2\pi(\vec{W}^T C\vec{W} + \sigma^2))\right] = \frac{1}{2}\log\left[1 + \frac{\vec{W}^T C\vec{W}}{\sigma^2}\right]$$

10 *Maximization of the mutual information*

The mutual information is maximum when the combination of weights maximizes the product $\vec{W}^T C\vec{W}$. Finding this optimal weight vector requires to determine $N-1$ unknowns, under the constraint $\|\vec{W}\|^2 = \sum_j w_j^2 = 1$.

To simplify this problem, it is relevant to use a change of coordinates in which the correlation matrix is diagonal. Since the matrix C is symmetrical and real, it can be diagonalized in an orthonormal basis.

Let v_j denote the coordinates of $\vec{W}$ in this new basis and λ_j the eigenvalues of C.
The product becomes : $\vec{W}^T C\vec{W} = \sum_j \lambda_j v_j^2$. The norm is unchanged : $\|\vec{W}\|^2 = \sum_j v_j^2 = 1$

The product is maximized by assigning a value of 1 to the coefficient v_i multiplied by the largest eigenvalue λ_i^*, and 0 to all other coefficients. This is equivalent to say that the vector $\vec{W}$ is aligned with the eigenvector of C associated with the largest eigenvalue. 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

(11) *Positivity*

$KL(p||q) = \int \mathrm{d}x\, p(x) \log \left[\frac{p(x)}{q(x)}\right] = -\int \mathrm{d}x\, p(x) \log \left[\frac{q(x)}{p(x)}\right]$
Using the inequality $\ln(x) \leq x - 1,\ \forall x > 0$:

$$\begin{aligned}\int \mathrm{d}x\ p(x) \log\left(\frac{q(x)}{p(x)}\right) &\leq \int \mathrm{d}x\ p(x)\left(\frac{q(x)}{p(x)} - 1\right)\\ &= \int \mathrm{d}x\ q(x) - \int \mathrm{d}x\ p(x)\\ &= 1 - 1 = 0\end{aligned}$$

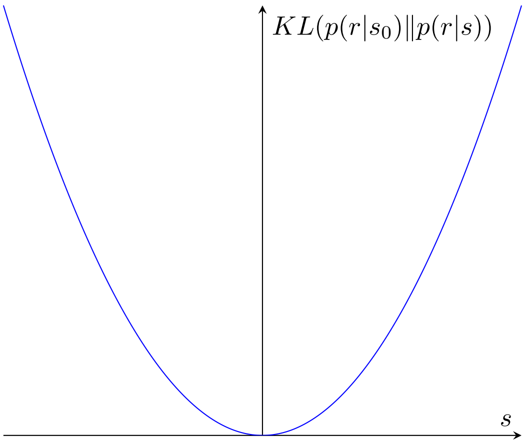
Therefore : $-KL(p||q) \leq 0 \implies KL(p||q) \geq 0$

Cancelation of the first derivative

The first derivative cancels at s_0 because $KL(p(r|s_0)||p(r|s))$ then cancells. Therefore, it reaches its minimum.

(12) *Sketch of the Kullbach-Liebler divergence*

Using the results above, the Kullback Liebler divergence is always positive, and cancels when both distributions are equal, i.e. for $s = s_0$.



(13) *Measure of local information*

The expression $F(s_0) = \frac{\partial^2 KL(p(r|s_0)||p(r|s))}{\partial s^2}\Big|_{s_0}$ requires to evaluate the second derivative at s_0 :

$$\begin{aligned}\frac{\partial^2 KL(p(r|s_0)||p(r|s))}{\partial s^2} &= \frac{\partial^2}{\partial s^2}\int p(r|s_0)\log\left(\frac{p(r|s_0)}{p(r|s)}\right) dr\\ &= \frac{\partial^2}{\partial s^2}\int p(r|s_0)\log(p(r|s_0))dr - \frac{\partial^2}{\partial s^2}\int p(r|s_0)\log(p(r|s))dr\\ &= 0 - \int p(r|s_0)\frac{\partial^2}{\partial s^2}\log(p(r|s))dr\end{aligned}$$

The first term cancels because it does not depend on s.

Conclusion : $F(s_0) = -\int p(r|s_0)\left[\frac{\partial^2}{\partial s^2}\log(p(r|s))\right]\Big|_{s_0} dr$

(14) *Interpretation*

The Fisher metric provides a measure of "how fast" both distributions separate out.

2.2 Variance of the locally optimal estimator

15 *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 :

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

16 *Equality between two formulas*

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

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

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

17 *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

18 *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.

19 *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 :

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

20 *Neuron with Gaussian noise*

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 :

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

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

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

21 *Two independent neurons*

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

$$\begin{gathered}
p(r_1, r_2|s) = p(r_1|s)p(r_2|s) \\
\log(p(r_1, r_2|s)) = \log(p(r_1|s)) + \log(p(r_2|s))
\end{gathered}$$

3 Bayesian inference

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

Because all the neurons are independent, the probability of a pattern is the product of the probabilities of observing each spike count for each neuron.

The probability of a spike count $k = n_i$ is given by the Poisson distribution with mean $\lambda = f_i(s)$: $\mathbb{P}(k) = \frac{\lambda^k}{k!} e^{-\lambda}$

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

(23) *Estimate of the stimulus*

An estimate of the stimulus can be built by 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.

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

Using previous results, the Fisher information is

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

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).

(25) *Probability distribution of the stimulus*

Using Baye's rule :

$$\begin{aligned} p(s|\{n_i\}) &= \frac{p(s)p(\{n_i\}|s)}{p(\{n_i\})} \\ &= \frac{p(s)}{p(\{n_i\})} \prod_{i=1}^{N} \frac{f_i(s)^{n_i}}{n_i!} \exp(-f_i(s)) \\ &= \frac{p(s)}{p(\{n_i\})} \left(\prod_{i=1}^{N} \frac{1}{n_i!} \right) \exp\left(-\sum_{i=1}^{N} f_i(s) \right) \exp\left(\sum_{i=1}^{N} n_i \log(f_i(s)) \right) \end{aligned}$$

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. Moreover, by assumption, the prior $p(s)$ is uniform, therefore it does not depend on s either. Thus :

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

(26) *Gaussian tuning curves*

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 parameters of the Gaussian can be deetermine by rewriting the quantity inside the exponential under the form of the gaussian distribution : $\frac{(s-\mathbb{E}(s))^2}{\mathbb{V}(s)}$. It involves finding the canonical form of this sum of squares :

$$\begin{aligned}
\sum_{i=1}^{N} n_i(s-s_i)^2 &= \left(\sum_{i=1}^{N} n_i\right) s^2 - 2\left(\sum_{i=1}^{N} n_i s_i\right) s + \sum_{i=1}^{N} n_i s_i^2 \\
&= \left(\sum_{i=1}^{N} n_i\right)\left[s^2 - 2\frac{\sum_{i=1}^{N} n_i s_i}{\sum_{i=1}^{N} n_i} s\right] + \sum_{i=1}^{N} n_i s_i^2 \\
&= \left(\sum_{i=1}^{N} n_i\right)\left(s - \frac{\sum_{i=1}^{N} n_i s_i}{\sum_{i=1}^{N} n_i}\right)^2 + \text{constant independent of } s
\end{aligned}$$

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

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

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.

27 *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.