THEORETICAL NEUROSCIENCE I

Lecture 16: Fisher information

Prof. Jochen Braun

Otto-von-Guericke-Universität Magdeburg,
Cognitive Biology Group

1 Statistical estimation in sensory systems

An animal's knowledge of the outside world usually derives from sensors that are noisy, biased, distorted, or corrupted in some other way (ganglion cells in the retina, hair cells in the cochlea, mechanoreceptors in the skin, etc). There are two reasons why animals nevertheless extract accurate information about the outside world.

(1) Most sensory systems rely on a large number of sensors (5M gangion cells in the retina, 5k inner hair cells in the cochlea, etc.), so that individual limitations can be overcome by averaging.

(2) The information from multiple sensors is combined in a "statistically efficient" way, using Bayesian principles.

This statistically efficient approach underlies the amazing sensory feats you see in nature movies and also makes possible recent technological advances such as wireless communication.

However, there is a limit to statistically efficient estimation. Given the number and properties of sensors, there is an upper bound to the accuracy and reliability that can be achieved ("Cramer-Rao bound").

1.1 Sensory estimation

To formalize sensory estimation, we introduce a stimulus value s a set of responses $\{x_i\}$ of N sensors, and a abstract function $T()$, which estimates the stimulus value on the basis of observed responses.

$$\text{stimulus } s \qquad \rightarrow \qquad \text{responses } \{\mathrm{x_i}\} \qquad \rightarrow \qquad s_{est} = T\left(\{x_i\}\right)$$

The information that sensors carry about s depends on their conditional response probability $p(x_i \,|\, s)$:

$$\text{conditional response probability} \qquad p(x_i \,|\, s)$$

Only with the 'prior knowledge' of $p(x_i|s)$ can one hope to interpret a response x_i. Accordingly, the estimation function $T()$ must be derived from the $p(x_i|s)$ of all N sensors:

$$p(x_i|s)\ \forall\, i \qquad \rightarrow \qquad s_{est} = T\left(\{x_i\}\right)$$

1.2 Quality of estimates

The quality of repeated estimates is measured in terms of the *bias* b_{est} and the *variance* σ^2_{est} of the estimated stimulus value s_{est}:

$$\text{Bias} \qquad b_{est}(s) = \langle s_{est} \rangle - s \qquad\qquad \text{unit}[s]$$

$$\text{Variance} \qquad \sigma^2_{est}(s) = \langle (s_{est} - \langle s_{est} \rangle)^2 \rangle \qquad\qquad \text{unit}[s^2]$$

Given these two quantities, we can determine the average squared estimation error

$$\langle (s_{est} - s)^2 \rangle = \langle (s_{est} - \langle s_{est} \rangle + b_{est})^2 \rangle = \sigma^2_{est} + b^2_{est}$$

1.3 Lower bounds for bias and varaince

In this lecture we are interested in *lower bounds* for the bias and variance of a stimulus estimate $T(\{x_i\})$.

The lower bound for bias is zero. In this case we speak of an *unbiased* estimate:

$$b_{est}(s) \geq 0$$

The lower bound for variance is a finite number, which will turn out to depend on the true stimulus value s and on the conditional response probabilities $p(\{x_i\}|s)$

$$\sigma^2_{est}(s) \geq \frac{1}{J_s(s)} > 0 \qquad\qquad J_s(s) = f\left[p\left(\{x_i\}|s\right)\right]$$

Here, $J_s(s)$ is the *Fisher Information* with respect to stimulus parameter s. The Fisher information reveals the highest accuracy (lowest variance) that can be achieved.

Fisher information is always **about** a particular stimulus value **s**.

Fisher information has units of

$$\frac{1}{s^2}$$

Fisher information gives smallest possible variance (standard deviation) of estimate:

$$\text{variance} \qquad \sigma^2_{est} \geq \frac{1}{J_s}$$

$$\text{standard deviation} \qquad \sigma_{est} \geq \sqrt{\frac{1}{J_s}}$$

$$s$$

$$s_{est}$$

$$\sqrt{\frac{1}{J_s}}$$

2 Example: Gaussian tuning for position

Consider a population of N neurons with Gaussian tuning for stimulus position s. Each neuron i prefers a stimulus position s_i and its average response decreases as the difference between actual position s and preferred position s_i increases.

2.1 Tuning curves

Specifically, we assume that the *average response* $r_i(s)$ of neuron i follows a Gaussian tuning curve:

$$r_i(c,s) = c \exp\left(-\frac{(s-s_i)^2}{2\sigma_s^2}\right) \qquad \frac{\delta r_i}{\delta c} = \frac{1}{c}\, r_i \qquad \frac{\delta r_i}{\delta s} = -\frac{s-s_i}{\sigma_s^2}\, r_i$$

where c is stimulus contrast and σ_s is the width of the tuning curve. We further assume

$$i = 1, 2, \ldots, 100 \qquad s_i = 10 \cdot i \qquad \sigma_s = 200 \qquad s_{true} = 500.0 \qquad c_{true} = 0.5$$

2.2 Response variability

The *actual response* x_i of neuron i will be different each time the stimulus is presented. Specifically, we assume a Gaussian distribution of constant width σ_x around the average response r_i. The conditional probability of observing an actual response x_i from neuron i, given a stimulus position $s = s_{true}$, is:

$$p(x_i|s) = \frac{1}{\sqrt{2\pi}\,\sigma_x} \exp\left[-\frac{(x_i - r_i)^2}{2\sigma_x^2}\right] \qquad \frac{\delta p(x_i|s)}{\delta r_i} = -\frac{r_i - x_i}{\sigma_x^2}\, p(x_i|s)$$

This distribution has the following expectation values

$$\langle x_i \rangle = r_i \qquad \langle x_i^2 \rangle = r_i^2 + \sigma_x^2$$

We assume $\sigma_x = 0.2$.

2.3 An estimation procedure (response-weighted average)

To estimate stimulus position from an observed population response, we weight the preferred position of each neuron with its response and form the average:

$$s'_{est} = \frac{\sum_i s_i\, x_i}{\sum_i x_i}$$

To estimate stimulus contrast, we compute the average response and multiply with an empirical factor

$$c'_{est} = \frac{2.03}{N} \sum_i x_i$$

2.4 Another estimation procedure (maximum likelihood)

2.4.1 Likelihood of population response

We now derive the probability of observing a set of responses x_1, x_2, …, x_N, given a stimulus position $s = s_{true}$. Assuming that neuronal responses vary independently, the joint probability of obtaining x_1, x_2, …, x_N is simply the product of the individual probabilities:

$$P(x_1, x_2, \ldots, x_N|s) = p(x_1|s) \cdot p(x_2|s) \cdot \ldots \cdot$$
$$p(x_N|s) = \prod_i p(x_i|s) = \frac{1}{(2\pi)^{N/2}\, \sigma_x^N} \prod_i \exp\left[-\frac{(x_i - r_i)^2}{2\sigma_x^2}\right]$$

Taking the logarithm, we obtain the likelihood of a set of responses x_1, x_2, …, x_N:

$$L(x_1, x_2, \ldots, x_N|s) = \ln P(x_1, x_2, \ldots, x_N|s)$$
$$= -\frac{N}{2} \ln(2\pi) - N \ln \sigma_x - \frac{1}{2\sigma_x^2} \sum_i [x_i - r_i]^2$$

2.4.2 Maximizing the likelihood of observed response

We can use the above equation to find the stimulus position s_{est} and stimulus contrast c_{est} that *maximizes the likelihood* of an observed set of responses x_1, x_2, …, x_N. In doing so, we may ignore terms and factors that do not depend on s, as they won't influence the position of the maximum. The expression that remains is

$$L(x_1, x_2, \ldots, x_N|s) \propto -\sum_i [x_i - r_i]^2 = -\sum_i \left[x_i - c\, \exp\left(-\frac{(s - s_i)^2}{2\sigma_s^2}\right)\right]^2$$

Compute this expression as a function of s and c. Use the Matlab functions *find* and *max* to locate the maximum. The values of s and c that maximize L are the desired estimates c_{est} and s_{est}.

Note that the computation reduces to a least-squares fit.

2.5 How good is the best estimate?

Given an estimation procedure, we can determine its quality experimentally. To this end, we need to stimulate repeatedly with known stimuli, observe the population response evoked by each, and compare estimated and true stimulus values.

$$s_{true}, c_{true} \quad \rightarrow \quad \{x_i\} \quad \rightarrow \quad s_{est} = T_s\left(\{x_i\}\right), \, c_{est} = T_c\left(\{x_i\}\right)$$

After a sufficient number of trials, we can determine the variances of our estimates

$$\sigma_s^2 = \frac{1}{N_{trial}} \sum_n \left[s_n^{est} - s_n^{true}\right]^2. \qquad \sigma_c^2 = \frac{1}{N_{trial}} \sum_n \left[c_n^{est} - c_n^{true}\right]^2$$

Alternatively, the Fisher information is theoretical way to directly determine how good the best possible estimate can be:

$$p(x_i|s)\,\forall\, i \quad \rightarrow \quad J_s, J_c \quad \rightarrow \quad \sigma_s^2 \geq \frac{1}{J_s}, \; \sigma_c^2 \geq \frac{1}{J_c}$$

3 Fisher information of a single neuron

We now derive the Fisher information of a single neuron. We are given an observed response x, the neuron's conditional response probability $p(x|s)$, and its derivative wrt s.

We want to compute the smallest possible variance of an estimate, σ_{est}^2. To make the connection, we assume that there exists an estimator function $s_{est} = T(x)$. We do *not* specify this function, but merely assume its existence.

$$\begin{array}{ll} x & p(x|s), \dfrac{\delta}{\delta s}\, p(x|s) \\ s_{est} = T(x) & \sigma_{est}^2 = \left\langle \left[T - \langle T \rangle\right]^2 \right\rangle \end{array}$$

3.1 Score function $V_s(x)$

To derive the fundamental limit mentioned above, we consider first a score function $V_s(x)$ which measures the proportional dependence of the distribution $p(x|s)$ on stimulus parameter s. We are particularly interested in the expectation values of this function and of its square:

$$V_s(x) \equiv \frac{\delta}{\delta s}\, \ln p(x|s) = \frac{1}{p(x|s)}\, \frac{\delta}{\delta s} p(x|s)$$

$$\langle V_s(x) \rangle = \int V_s(x)\, p(x|s)\, dx = \int \frac{\delta}{\delta s} p(x|s)\, dx = \\ \frac{\delta}{\delta s} \int p(x|s)\, dx = \frac{\delta}{\delta s}\, 1 = 0$$

$$\langle [V_s(x) - \langle V_s(x) \rangle]^2 \rangle = \langle V_s^2(x) \rangle$$

Example:

For one of our position-tuned neurons, we have a score function about the mean response r_i

$$V_{r_i}(x_i) \equiv \frac{\delta}{\delta r_i} \ln p(x_i|r_i) = \frac{1}{p(x_i|r_i)} \frac{\delta}{\delta r_i} p(x_i|r_i) = -\frac{r_i - x_i}{\sigma_x^2}$$

a score function about the stimulus position s

$$V_s(x_i) \equiv \frac{\delta}{\delta s} \ln p(x_i|s) = V_{r_i} \frac{\delta r_i}{\delta s} = \frac{\delta}{\delta r_i} \ln p(x_i|r_i) \frac{\delta r_i}{\delta s} r_i(s) = \frac{r_i - x_i}{\sigma_x^2} \frac{s - s_i}{\sigma_s^2} r_i$$

and a score function about the stimulus contrast c

$$V_c(x_i) \equiv \frac{\delta}{\delta c} \ln p(x_i|c) = V_{r_i} \frac{\delta r_i}{\delta c} = \frac{\delta}{\delta r_i} \ln p(x_i|r_i) \frac{\delta r_i}{\delta c} r_i(c) = -\frac{r_i - x_i}{\sigma_x^2} \frac{r_i}{c}$$

if we recall that

$$\frac{\delta p(x_i|s)}{\delta r_i} = -\frac{r_i - x_i}{\sigma_x^2} p(x_i|s) \qquad \frac{\delta r_i}{\delta s} = -\frac{s - s_i}{\sigma_s^2} r_i \qquad \frac{\delta r_i}{\delta c} = -\frac{1}{c} r_i$$

3.2 Fisher information J_s

The *Fisher information* is defined as the expectation value of the square of the score function.

$$\textit{Fisher information} \qquad J_s \equiv \langle V_s^2(x) \rangle$$

$$J_s \equiv \int V_s^2(x)\, p(x|s)\, dx$$

It is not an information in the Shannon sense but relates to the quality of estimators (as mentioned above).

Example:

For one of our position-tuned neurons, the Fisher information about mean response r_i, about stimulus position s and about stimulus contrast c follows from the respective score functions:

$$J_{r_i} = \int V_{r_i}^2(x_i)\, p(x_i|r_i)\, dx_i = \int \left[\tfrac{r_i - x_i}{\sigma_x^2}\right]^2 p(x_i|r_i)\, dx_i = \tfrac{1}{\sigma_x}^2 \int \left[r_i^2 - 2\, r_i\, x_i + x_i^2\right] p(x_i|r_i)\, dx_i =$$

$$= \frac{1}{\sigma_x^4} \left[r_i^2 - 2\, r_i\, \langle x_i \rangle + \langle x_i^2 \rangle\right] = \frac{1}{\sigma_x^2}$$

$$J_s = \int V_s^2(x_i)\, p(x_i|r_i)\, dx_i = \int V_{r_i}^2(x_i) \left(\tfrac{\delta r_i}{\delta s}\right)^2 p(x_i|r_i)\, dx_i = \left(\tfrac{\delta r_i}{\delta s}\right)^2 \int V_{r_i}^2(x_i)\, p(x_i|r_i)\, dx_i =$$

$$= \left(\frac{\delta r_i}{\delta s}\right)^2 J_{r_i} = \frac{(s - s_i)^2}{\sigma_s^4}\, r_i^2\, \frac{1}{\sigma_x^2}$$

$$J_c = \int V_c^2(x_i)\, p(x_i|r_i)\, dx_i = \left(\tfrac{\delta r_i}{\delta c}\right)^2 J_{r_i} =$$

$$= \frac{1}{c^2}\, r_i^2\, \frac{1}{\sigma_x^2}$$

if we recall that $\langle x_i \rangle = r_i$ and $\langle x_i^2 \rangle = r_i^2 + \sigma_x^2$.

3.3 Cramer-Rao bound

We now derive the Cramer-Rao bound, which relates Fisher information to the variance of any estimator. We assume an ‘unbiased’ estimator with $\sigma_b^2 = 0$. Our starting point is the Cauchy-Schwartz inequality for integrals:

$$\langle V_s \rangle = 0 \qquad J_s = \langle V_s^2 \rangle = \langle (V_s - \langle V_s \rangle)^2 \rangle \qquad \sigma_s^2 = \langle (T_s - \langle T_s \rangle)^2 \rangle$$

$$\langle (V_s - \langle V_s \rangle)\, (T_s - \langle T_s \rangle) \rangle^2 \leq \langle (V_s - \langle V_s \rangle)^2 \rangle\, \langle (T_s - \langle T_s \rangle)^2 \rangle$$

$$(\mathrm{Cauchy - Schwartz\ inequality})$$

$$\left[\langle V_s\, T_s \rangle - \langle V_s \rangle\, \langle T_s \rangle\right]^2 \leq J_s\, \sigma_s^2$$

$$\langle V_s\, T_s \rangle^2 \leq J_s\, \sigma_s^2$$

$$\langle V_s\, T_s \rangle^2 \leq J_s\, \sigma_s^2$$

$$\langle V_s\, T_s \rangle = \int \frac{1}{p(x|s)} \frac{\delta}{\delta s} p(x|s)\, T_s(x)\, p(x|s)\, dx = \int \frac{\delta}{\delta s} p(x|s)\, T_s(x)\, dx =$$

$$= \frac{\delta}{\delta s} \int p(x|s)\, T_s(x)\, dx = \frac{\delta}{\delta s} \langle T_s(x) \rangle = \frac{\delta s}{\delta s} = 1$$

$$\Rightarrow \qquad \sigma_s^2 \geq \frac{1}{J(s)}$$

Note that the last step assumes an 'unbiased' estimator for which $\langle T_s(s) \rangle = s$

Thus, the inverse of the Fisher information is a lower bound for the variance of any unbiased estimator.

Other example:

Consider a population of neurons tuned to log spatial and log temporal frequency of a visual stimulus.

From the tuning function and response distribution we can compute the respective Fisher informations $J_{\ln \omega_r}$ and $J_{\ln \omega_t}$.

Since log temporal and log spatial frequency together determine the speed of such a stimulus, we can also compute Fisher information about log speed:

$$v = \omega_t / \omega_r$$

$$\ln \omega_t = \ln \omega_r + \ln v \qquad \qquad \ln \omega_r = \ln \omega_t - \ln v$$

$$\frac{\delta \ln \omega_t}{\delta \ln v} = 1 \qquad \qquad \frac{\delta \ln \omega_r}{\delta \ln v} = -1$$

$$J_{\ln v} = \left[\frac{\delta \ln \omega_t}{\delta \ln v} \right]^2 J_{\ln \omega_t} + \left[\frac{\delta \ln \omega_r}{\delta \ln v} \right]^2 J_{\ln \omega_r} = J_{\omega_t} + J_{\omega_r}$$

Accordingly, the information for stimulus velocity comprises two components, one reflecting the filter tuning to spatial frequency and the other the tuning to temporal frequency.

4 Fisher information of multiple neurons

4.1 Additivity of Fisher information

Fisher information is useful because it is *additive*. In other words, the FI of multiple neurons is simply the sum of the FI of the individual neurons, assuming independence.

$$\text{n independent neurons :} \qquad p(x_1, x_2, \ldots, x_n|s) = \prod_i p(x_i|s)$$

Score functions are additive:

$$\begin{aligned} V_s(x_1, x_2, \ldots, x_n) &= \frac{\delta}{\delta s} \ln p(x_1, x_2, \ldots, x_n|s) \\ &= \sum_i \frac{\delta}{\delta s} \ln p(x_i|s) = \sum_i V_s(x_i) \end{aligned}$$

So how can the expectation value of the squared score function also be additive???

Fisher informations are additive:

$$J_s^{all} = \langle V_s^2(x_1, x_2, \ldots, x_n) \rangle = \langle \left(\sum_i V_s(x_i) \right)^2 \rangle =$$

$$= \langle \sum_i V_s^2(x_i) \, + \sum_{i \neq j} V_s(x_i)\, V_s(x_j) \rangle = \sum_i \langle V_s^2(x_i) \rangle + \sum_{i \neq j} \langle V_s(x_i)\, V_s(x_j) \rangle =$$

$$= \sum_i \langle V_s^2(x_i) \rangle + \sum_{i \neq j} \langle V_s(x_i) \rangle \, \langle V_s(x_j) \rangle =$$

$$= \sum_i \langle V_s^2(x_i) \rangle = \sum_i J_s^i$$

Thus, because of the assumption of independence, the combined Fisher information of multiple neural responses $\{x_i\}$ is simply the sum of the Fisher informations of the individual responses x_i.

$$J_s^{all} = \sum_i J_s^i$$

Because of the inverse relation between Fisher information and variance, we can equivalently say that the inverse variance of the combined estimate $T(\{x_i\})$ equals the sum of the inverse variances of the individual estimates $t_i(x_i)$.

$$J_s^{all} = \frac{1}{(\sigma_s^{all})^2} = \sum_i J_i(s) = \sum_i \frac{1}{(\sigma_s^i)^2}$$

5 Linear combination of estimates

With the help of the Fisher information approach, we have seen that inverse variances are additive when individual estimates are combined optimally.

This result can also be derived more simply, by considering the optimal linear combination of individual estimates.

Specifically, one may construct a population estimate $T_{all}(\{x_i\})$ from a linear combination of individual unbiased estimates $t_i(x_i)$.

To simplify, we assume $s_{true} = 0$.

$$\langle t_i \rangle = s_{true} = 0 \qquad \sigma_i^2 = \langle t_i^2 \rangle - \langle t_i \rangle^2 = \langle t_i^2 \rangle$$

$$T_{all} = \sum_i \alpha_i \, t_i \qquad \sum_i \alpha_i = 1$$

Assuming linear weights α_i, we derive the variance of the population estimate σ_{all}^2 from the variances of the individual estimates σ_i^2:

$$\langle T_{all} \rangle = \langle \sum_i \alpha_i \, t_i \rangle = \sum_i \alpha_i \, \langle t_i \rangle = 0$$

$$\sigma_{all}^2 = \langle T_{all}^2 \rangle - \langle T_{all} \rangle^2 = \langle T_{all}^2 \rangle = \langle \sum_i \alpha_i \, t_i \, \sum_j \alpha_j \, t_j \rangle =$$

$$= \sum_i \langle \alpha_i^2 \, t_i^2 \rangle + \sum_{j \neq i} \langle \alpha_i \, t_i \rangle \, \langle \alpha_j \, t_j \rangle =$$

$$= \sum_i \alpha_i^2 \, \langle t_i^2 \rangle + \sum_{j \neq i} \alpha_i \, \alpha_j \, \langle t_i \rangle \, \langle t_j \rangle =$$

$$= \sum_i \alpha_i \, \langle t_i^2 \rangle = \sum_i \alpha_i^2 \, \sigma_i^2$$

Our next task is to choose the weights α_i such as to minimize the variance σ_{all}^2 of the population estimate:

$$0 = \frac{\delta}{\delta \, \alpha_i} \, \sigma_{all}^2 = \frac{\delta}{\delta \, \alpha_i} \left[\sum_{j \neq N} \alpha_j^2 \, \sigma_j^2 + \alpha_N^2 \, \sigma_N^2 \right]$$
$$= \frac{\delta}{\delta \, \alpha_i} \left[\sum_{j \neq N} \alpha_j^2 \, \sigma_j^2 + \left(1 - \sum_{j \neq N} \alpha_j \right)^2 \sigma_N^2 \right]$$

$$0 = 2\,\alpha_i\,\sigma_i^2 - 2\left(1 - \sum_{j \neq N} \alpha_j\right)\sigma_N^2 \qquad \Leftrightarrow \qquad \alpha_i\,\sigma_i^2 = \alpha_N\,\sigma_N^2$$

$$1 = \sum_i \alpha_i = \alpha_N\,\sigma_N^2 \sum_i \frac{1}{\sigma_i^2} \qquad \Leftrightarrow \qquad \alpha_N = \frac{\frac{1}{\sigma_N^2}}{\sum_i \frac{1}{\sigma_i^2}}$$

In words, the optimal linear weight α_i is inversely proportional to the individual variance σ_i^2.

Having found the optimal weights, we can calculate the optimal variance of the population estimate:

$$\sigma_{all}^2 = \sum_i \alpha_i^2\,\sigma_i^2, \qquad \alpha_i = \frac{\frac{1}{\sigma_i^2}}{\sum_k \frac{1}{\sigma_k^2}}$$

$$\sigma_{all}^2 = \sum_i \frac{\frac{1}{\sigma_i^4}\,\sigma_i^2}{\left(\sum_k \frac{1}{\sigma_k^2}\right)^2} = \frac{\sum_i \frac{1}{\sigma_i^2}}{\left(\sum_k \frac{1}{\sigma_k^2}\right)^2} = \frac{1}{\sum_i \frac{1}{\sigma_i^2}}$$

$$\frac{1}{\sigma_{all}^2} = \sum_i \frac{1}{\sigma_i^2}$$

In words, for an optimal linear combination of individual estimates, the inverse variance of the combined estimate equals the sum of the inverse variances of the individual estimates.

6 Fisher information and neuron properties

With the help of Fisher information, we can determine how the stimulus information encoded by a population depends on individual neuron properties.

To illustrate this point, we now return to our example of position-tuned neurons. Recall the individual Fisher informations about stimulus position s and stimulus contrast c, which set a lower bound for the variances of any individual estimates:

$$J_s^i = \frac{(s_{true} - s_i)^2}{\sigma_s^4} \, r_i^2 \, \frac{1}{\sigma_x^2} \qquad \geq \qquad \frac{1}{\sigma_{s\ est\ i}^2}$$

$$J_c^i = \frac{1}{c_{true}^2} \, r_i^2 \, \frac{1}{\sigma_x^2} \qquad \geq \qquad \frac{1}{\sigma_{c\ est\ i}^2}$$

For the entire population, the combined Fisher information about stimulus position s and about stimulus contrast c are

$$J_s^{all} = \sum_i J_s^i = \frac{c_{true}^2}{\sigma_x^2 \, \sigma_s^4} \sum_i (s_{true} - s_i)^2 \, \exp\left[-\frac{(s_{true} - s_i)^2}{\sigma_s^2}\right] \geq \frac{1}{\sigma_{s\ est\ all}^2}$$

$$J_c^{all} = \sum_i J_c^i = \frac{1}{\sigma_x^2} \sum_i \exp\left[-\frac{(s_{true} - s_i)^2}{\sigma_s^2}\right] \qquad \geq \qquad \frac{1}{\sigma_{c\ est\ all}^2}$$

if we recall that

$$r_i = c_{true} \, \exp\left[-\frac{(s_{true} - s_i)^2}{2\sigma_s^2}\right]$$

$$\sigma_{s\ est\ all} \geq \frac{1}{\sqrt{J_s^{all}}}$$

$$\sigma_{c\ est\ all} \geq \frac{1}{\sqrt{J_c^{all}}}$$

What does this result tell us? The combined Fisher information about position s increases with the number of neurons, decreases with the response variance σ_x, and reaches a maximum for a particular tuning width σ_s:

$$J_s^{all} = \frac{c_{true}^2}{\sigma_x^2 \, \sigma_s^4} \sum_i (s_{true} - s_i)^2 \, \exp\left[-\frac{(s_{true} - s_i)^2}{\sigma_s^2}\right] \qquad \geq \qquad \frac{1}{\sigma_{s\ est\ all}^2}$$

The combined Fisher information about contrast c increases with the number of neurons, decreases with response variance σ_x, and increases with tuning width σ_s:

$$J_c^{all} = \sum_i J_c^i = \frac{1}{\sigma_x^2} \sum_i \exp\left[-\frac{(s_{true} - s_i)^2}{\sigma_s^2}\right] \qquad \geq \qquad \frac{1}{\sigma^2_{c\ est\ all}}$$

7 Summary

Neuronal responses to a stimulus s are characterized by the conditional response distribution $p(x|s)$.

Given $p(x|s)$ and an observed response x, we can estimate the value of s.

Many estimation procedures are possible, for example maximum likelihood estimation (pick the s that maximizes $p(x|s)$).

All estimation procedures are limited by the Cramer-Rao bound, in that their variance cannot be smaller than this bound.

Given $p(x|s)$, we can compute Fisher information and the Cramer-Rao bound:

$$\text{FI}: \quad J_s \qquad\qquad \text{CR bound}: \quad \sigma_s \geq \frac{1}{\sqrt{J_s}}$$

Thus, $p(x|s)$ limits the information about s than can be obtained from x.

In practise, FI is useful because it is additive and thus can be easily calculated for a population of neurons with different $p_i(x_i|s)$.

Thus, it reveals the respective influence of neuron number, response variance, and tuning width on the informativeness of a population response.

It also reveals which members of a neuronal population are particularly informative about a given stimulus property.

8 Fisher information and entropy

For the sake of completeness and for future reference, we note a relation between the Fisher information and the relative entropy, also known as the Kullback-Leibler distance, between two distributions $p_s(x)$ and $p_\theta(x)$:

$$D(p_s||p_\theta) = \int p_s(x) \ln \frac{p_s(x)}{p_\theta(x)}\, dx$$

We now consider the limit of $\theta \to s$, in which the two distributions become identical.

$$D(p_s||p_\theta)|_{\theta=s} = 0$$

$$\frac{\delta}{\delta\theta}\, D(p_s||p_\theta)|_{\theta=s} = \lim_{\Delta s \to 0} \frac{1}{\Delta s} \int p_s(x) \ln \frac{p_s(x)}{p_{s+\Delta s}(x)}\, dx = 0$$

$$\frac{\delta^2}{\delta\theta^2}\, D(p_s||p_\theta)|_{\theta=s} = \lim_{\Delta s \to 0} \frac{1}{\Delta s^2} \int p_s(x) \ln \frac{p_s(x)}{p_{s+\Delta s}(x)}\, dx = J(s)$$

Accordingly, the Fisher information is the second derivative of the relative entropy at a particular value of s.

8.1 Examples of response distributions and score functions

Exponential distribution with mean s:

$$p_s(x) = \frac{1}{s} \exp\left(-\frac{x}{s}\right) \qquad \int_0^\infty p_s(x)\, dx = 1 \qquad \langle x \rangle = s \qquad \langle x^2 \rangle = 2\, s^2$$

$$V_s(x) = \frac{1}{p_s(x)} \frac{\delta}{\delta s} p_s(x) = \frac{x - s}{s^2}$$

$$J_s = \int_0^\infty \left(\frac{x-s}{s^2}\right)^2 p_s(x)\, dx = \frac{1}{s^4} \left[\langle x^2 \rangle - 2s \langle x \rangle + s^2\right] = \frac{1}{s^2}$$

Gaussian distribution with mean μ:

$$p_\mu(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right) \qquad \int_{-\infty}^\infty p_\mu(x)\, dx = 1 \qquad \langle x \rangle = \mu \qquad \langle x^2 \rangle = 1 + \mu^2$$

$$V_\mu(x) = \frac{1}{p_s(x)} \frac{\delta}{\delta \mu} p_\mu(x) = \mu - x$$

$$J_\mu = \int_{-\infty}^\infty (\mu - x)^2\, p_\mu(x)\, dx = \mu^2 - 2\mu \langle x \rangle + \langle x^2 \rangle = 1$$

Gaussian distribution with variance σ^2:

$$p_\sigma(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{x^2}{2\,\sigma^2}\right) \qquad \int_{-\infty}^\infty p_\sigma(x)\, dx = 1 \qquad \langle x^2 \rangle = \sigma^2 \qquad \langle x^4 \rangle = 3\sigma^4$$

$$V_\sigma(x) = \frac{1}{p_\sigma(x)} \frac{\delta}{\delta \sigma} p_\sigma(x) = \frac{x^2 - \sigma^2}{\sigma^3}$$

$$J_\sigma = \int_{-\infty}^\infty \left(\frac{x^2 - \sigma^2}{\sigma^3}\right)^2 p_\sigma(x)\, dx = \frac{1}{\sigma^3} \left[\langle x^4 \rangle - 2\sigma^2 \langle x^2 \rangle + \sigma^4\right] = \frac{2}{\sigma^2}$$

9 Neurons with Gaussian tuning and realistic noise

9.1 Tuning function

To illustrate the usefulness of Fisher information, we now consider the a population of neurons tuned to some stimulus parameter θ (e.g., orientation). Note that the development is completely general and independent of the physical interpretation of θ. In other words, θ could be stimulus contrast, stimulus frequency, stimulus speed, or any other physical aspect. We assume the following average response and derivative:

$$\text{Tuning function} \qquad r(\theta) = \exp\left(-\frac{(\theta-\Theta)^2}{2s_\theta^2}\right)$$

$$\text{Derivative} \qquad \frac{\delta}{\delta\theta} r(\theta) = -\frac{\theta-\Theta}{s_\theta^2}\exp\left(-\frac{(\theta-\Theta)^2}{2s_\theta^2}\right) = -\frac{\theta-\Theta}{s_\theta^2}\, r(\theta)$$

where Θ is the preferred value of θ and s_θ is the Gaussian tuning width. The actual response $x(\theta)$ on any given trial depends on the average response $r(\theta)$ and on the distribution of noise. We make realistic assumptions about the noise in order to avoid singularities (especially when the mean response $r \to 0$).

9.2 Realistic noise

Specifically, we assume that variance σ_η^2 of the noise η includes a constant background component plus a second component that is proportional to the mean response: $\sigma_\eta^2 = \beta + \alpha\, r$. The conditional distribution of responses $p(x|r)$, given an average response $r(\theta)$:

$$x = r + \eta \qquad p(\eta)\, d\eta = \frac{1}{\sqrt{2\pi(\alpha\, r+\beta)}} \exp\left(-\frac{\eta^2}{2\alpha\, r + 2\beta}\right) d\eta$$

$$p(x|r)\, dx = \frac{1}{\sqrt{2\pi(\alpha\, r+\beta)}} \exp\left(-\frac{(x-r)^2}{2\alpha\, r + 2\beta}\right) dx$$

Both the mean response and the distribution of actual responses depend (indirectly) on stimulus parameters θ. The rate with which the distribution changes with a stimulus parameter θ is of the first interest, as it is this rate of change than provides information about θ.

$$\tfrac{\delta}{\delta s}\, p(x|r) = \tfrac{1}{\sqrt{2\pi(\alpha\, r+\beta)}} \left[-\tfrac{\alpha}{2(\alpha\, r+\beta)} - \tfrac{(r-x)}{\alpha\, r+\beta} + \tfrac{\alpha(x-r)^2}{2(\alpha r+\beta)^2}\right] \exp\left(-\tfrac{(x-r)^2}{2\alpha\, r+2\beta}\right) \tfrac{\delta}{\delta s}\, r(s) =$$

$$= p(x|r)\, N(x,r)\, \tfrac{\delta}{\delta s}\, r(s)$$

$$\text{where} \qquad N(x,r) = \tfrac{1}{\alpha r+\beta}\left[-\tfrac{\alpha}{2} + x - r + \tfrac{\alpha(x-r)^2}{2(\alpha r+\beta)}\right]$$

9.3 Score function and Fisher information for r

Before calculating the Fisher information with respect to stimulus parameter θ, we take the intermediate step of calculating Fisher information with respect to the *average response* $r(\theta)$. After all, it is only through $r(\theta)$ that our conditional response distribution depends on θ:

The score function for $V_r(x)$ with respect to r is

$$V_r(x) = \frac{1}{p(x|r)} \frac{\delta}{\delta r} p(x|r) = N(x,r) = \frac{1}{\alpha r + \beta} \left[-\frac{\alpha}{2} + x - r + \frac{\alpha (x-r)^2}{2(\alpha r + \beta)} \right]$$

and the Fisher information is

$$J_r \equiv \int p(x|r)\, N^2(x,r) dx = \left[\frac{1}{\alpha r + \beta} \right]^2 \left[\alpha r + \beta + \frac{\alpha^2}{2} \right]$$

$$J_r = \int p(x|r)\ [N(x,r)]^2\, dx =$$

$$\frac{1}{\sqrt{2\pi(\alpha\, r+\beta)}} \left[\tfrac{1}{\alpha r+\beta}\right]^2 \int \left[\tfrac{\alpha(x-r)^2}{2(\alpha\, r+\beta)} + (x-r) - \tfrac{\alpha}{2}\right]^2 \exp\left(-\tfrac{(x-r)^2}{2(\alpha r+\beta)}\right)\, dx =$$

$$= \frac{1}{\sqrt{2\pi(\alpha\, r+\beta)}} \left[\tfrac{1}{\alpha r+\beta}\right]^2 \int \left[\tfrac{\alpha^2(x-r)^4}{4(\alpha r+\beta)^2} + (x-r)^2 + \tfrac{\alpha^2}{4} + \tfrac{\alpha(x-r)^3}{\alpha\, r+\beta} - \tfrac{\alpha^2(x-r)^2}{2(\alpha\, r+\beta)} - \alpha(x-r)\right] e^{-\frac{(x-r)^2}{2(\alpha\, r+\beta)}}\, dx =$$

$$u = \tfrac{x-r}{\sqrt{\alpha r+\beta}} \qquad\qquad du = \tfrac{dx}{\sqrt{\alpha r+\beta}}$$

$$= \left[\tfrac{1}{\alpha r+\beta}\right]^2 \tfrac{1}{\sqrt{2\pi}} \left[\tfrac{\alpha^2}{4} \int u^4\, e^{-\frac{1}{2}u^2}\, du + (\alpha r+\beta) \int u^2\, e^{-\frac{1}{2}u^2}\, du + \tfrac{\alpha^2}{4} \int e^{-\frac{1}{2}u^2}\, du - \tfrac{\alpha^2}{2} \int u^2\, e^{-\frac{1}{2}u^2}\, du\right] =$$

$$= \left[\tfrac{1}{\alpha r+\beta}\right]^2 [\tfrac{3\alpha^2}{4} + (\alpha r+\beta) + \tfrac{\alpha^2}{4} - \tfrac{\alpha^2}{2}] = \left[\tfrac{1}{\alpha r+\beta}\right]^2 \left[\alpha r+\beta+\tfrac{\alpha^2}{2}\right]$$

9.4 Score function and Fisher information for θ

Now we take the final step and calculate the score function $V_\theta(x)$ and Fisher information J_θ with respect to θ:

$$V_\theta(x) = \frac{1}{p(x|\theta)} \frac{\delta}{\delta\theta} p(x|\theta) = N(x,r)\, \frac{\delta r(\theta)}{\delta\theta}$$
$$= \frac{1}{\alpha r + \beta} \left[-\frac{\alpha}{2} + x - r + \frac{\alpha(x-r)^2}{2(\alpha r+\beta)} \right] \frac{\delta r(\theta)}{\delta\theta}$$

$$J_\theta = \int p(x|r)\, N^2(x,r) \left(\frac{\delta r(\theta)}{\delta\theta}\right)^2 dx =$$

$$= \left[\frac{1}{\alpha r+\beta}\right]^2 \left[\alpha r + \beta + \frac{\alpha^2}{2}\right] \left(\frac{\delta r(\theta)}{\delta\theta}\right)^2 =$$

$$= \left[\frac{1}{\alpha r(\theta)+\beta}\right]^2 \left[\alpha r(\theta) + \beta + \frac{\alpha^2}{2}\right] \frac{(\theta-\Theta)^2}{s_\theta^4}\, r^2(\theta)$$

where

$$r(\theta) = \exp\left(-\frac{(\theta-\Theta)^2}{2 s_\theta^2}\right)$$

9.5 Conclusion

How accurately we may estimate θ on the basis of a population response $\{x_i\}$ and the associated conditional response probabilities follows from the total Fisher information. The variance of any estimate of $\theta = T(\{x_i\})$ is bounded by $\sigma_T^2 \geq 1/J_\theta(\{r_i\})$.

$$r_i = f(\theta_t, \Theta^{(i)}) = \frac{1}{\sqrt{2\pi} s_\theta} \exp \frac{-\left[\ln\theta - \ln\Theta^{(i)}\right]^2}{2 s_\theta^2}$$

$$J_\theta(\{r_i\}) = \sum_i J_\theta(r_i) = \sum_i \left[\frac{1}{\alpha r_i + \beta}\right]^2 \left[\alpha r_i + \beta + \frac{\alpha^2}{2}\right] \left(\frac{\delta r_i}{\delta\theta}\right)^2 =$$
$$= \sum_i \frac{1}{\alpha^2} \left[\frac{\alpha r_i}{\alpha r_i + \beta}\right]^2 \left[\alpha r_i + \beta + \frac{\alpha^2}{2}\right] \left[\frac{\theta - \Theta^{(i)}}{s_\theta^2}\right]^2$$

Feeling reckless, we omit dark noise by setting $\beta = 0$ and obtain

$$J_\theta(\{r_i\}) = \sum_i J_\theta(r_i) = \sum_i \left[\frac{\delta r_i/\delta\theta}{r_i}\right]^2 \left[\frac{r_i}{\alpha} + \frac{1}{2}\right] =$$
$$= \sum_i \left[\frac{\theta - \Theta^{(i)}}{s_\theta^2}\right]^2 \left[\frac{1}{\sqrt{2\pi}\, s_\theta\, \alpha} \exp \frac{-\left[\ln\theta - \ln\Theta^{(i)}\right]^2}{2 s_\theta^2} + \frac{1}{2}\right]^2$$

In this case, the Fisher information grows without bound as $|\theta - \Theta| \to \infty$, due to "information" contributed by filters with $r_i \approx 0$. Now we know why we cannot neglect dark noise!

10 Multiple stimulus dependencies

10.1 Two related stimulus variables

Consider a case in which the mean response r depends on a distribution parameter s, which in turn depends on stimulus variable s, $s = s(s)$. How are the associated Fisher informations J_s and J_s related in this case?

The answer follows from the definition of the Fisher information and the chain rule of derivatives:

$$J_s \equiv \int p(x|r) \, \left[\frac{\delta p(x|r)/\delta s}{p(x|r)} \right]^2 dx \qquad\qquad J_s \equiv \int p(x|r) \, \left[\frac{\delta p(x|r)/\delta s}{p(x|r)} \right]^2 dx$$

$$J_s = \int p(x|r) \, \left[\frac{\delta p(x|r)/\delta s}{p(x|r)} \right]^2 \, \left(\frac{\delta s}{\delta s} \right)^2 \, dx = \left(\frac{\delta s}{\delta s} \right)^2 \int p(x|r) \, \left[\frac{\delta p(x|r)/\delta s}{p(x|r)} \right]^2 \, dx = \left(\frac{\delta s}{\delta s} \right)^2 \, J_s$$

Accordingly, the Fisher information with respect to related variables differ only by a multiplicative factor, namely, the square of the "inner derivative". Due to the additivity of Fisher information, this results holds not only for a single neuron but equally for an entire population of neurons.

10.2 Three related stimulus variables

Consider a case in which the mean response r depends on two stimulus variables u and v that are related to a third variable s by $f(s,u,v)=0$. Can we define a Fisher information for J_s and relate it to the Fisher informations J_u and J_v? At every u,v, an infinitesimal increment ds entails infinitesimal increments du and dv, each of which changes $p(x|r)$ by an infinitesimal amount dp.

$$dp(x|r) = \left.\frac{\delta p(x|r)}{\delta u}\right|_{u,v} du + \left.\frac{\delta p(x|r)}{\delta v}\right|_{u,v} dv \qquad du = \left.\frac{\delta u}{\delta s}\right|_{u,v} ds \qquad dv = \left.\frac{\delta v}{\delta s}\right|_{u,v} ds$$

$$\frac{dp(x|r)}{ds} = \left.\frac{\delta p(x|r)}{\delta u}\right|_{u,v} \left.\frac{\delta u}{\delta s}\right|_{u,v} + \left.\frac{\delta p(x|r)}{\delta v}\right|_{u,v} \left.\frac{\delta v}{\delta s}\right|_{u,v}$$

$$J_s \equiv \int p(x|r)\left[\frac{1}{p(x|r)}\frac{\delta p(x|r)}{\delta s}\right]^2 dx \qquad J_u \equiv \int p(x|r)\left[\frac{1}{p(x|r)}\frac{\delta p(x|r)}{\delta u}\right]^2 dx \qquad J_v \equiv \int p(x|r)\left[\frac{1}{p(x|r)}\frac{\delta p(x|r)}{\delta v}\right]^2 dx$$

$$J_s = \int \frac{1}{p(x|r)} \left[\frac{\delta p(x|r)}{\delta u}\frac{\delta u}{\delta s}\right]^2 dx + 2\int \frac{1}{p(x|r)} \left[\frac{\delta p(x|r)}{\delta u}\frac{\delta v}{\delta s}\frac{\delta p(x|r)}{\delta v}\frac{\delta v}{\delta s}\right] dx + \int \frac{1}{p(x|r)} \left[\frac{\delta p(x|r)}{\delta v}\frac{\delta v}{\delta s}\right]^2 dx =$$

$$= \left[\frac{du}{ds}\right]^2 J_u + \left[\frac{du}{ds}\right]^2 J_v + 2\left[\frac{du}{ds}\right]\left[\frac{dv}{ds}\right]^2 \frac{\delta^2}{\delta u \delta v}\int p(x|r)\,dx = \left[\frac{du}{ds}\right]^2 J_u + \left[\frac{dv}{ds}\right]^2 J_v$$

Accordingly, the Fisher information for the third variable s is the sum of the Fisher information for u and v, weighted by the squre of the respective "inner derivatives".

11 Fisher information for 2D log Gaussian tuning

As a penultimate example, we consider a population with 2D tuning for log spatial frequency $\ln \omega_r$ and log temporal frequency $\ln \omega_t$. This is a good approximation of motion-selective neurons in area V1. We specify the tuning functions and compute the Fisher information for $\ln \omega_r$ and $\ln \omega_t$ with the same realistic noise assumption as before ($\sigma_\eta^2 = \alpha\, r + \beta$).

$$r_i = f(\omega_t, \omega_r, \Omega_t^{(i)}, \Omega_r^{(i)}) = \exp \frac{-\left(\ln \omega_r - \ln \Omega_r^{(i)}\right)^2}{2 s_r^2} \; \exp \frac{-\left(\ln \omega_t - \ln \Omega_t^{(i)}\right)^2}{2 s_t^2}$$

where $\Omega_r^{(i)}$ and $\Omega_t^{(i)}$ are the preferred frequencies and s_r and s_t the log Gaussian tuning widths.

$$J_{\ln \omega_t}(\{r_i\}) = \sum_i J_{\ln \omega_t}(r_i) = \sum_i \left[\frac{1}{\alpha r_i + \beta}\right]^2 \left[\alpha r_i + \beta + \frac{\alpha^2}{2}\right] \left(\frac{\delta r_i}{\delta \ln \omega_t}\right)^2 =$$
$$= \sum_i \left[\frac{r_i}{\alpha r_i + \beta}\right]^2 \left[\alpha r_i + \beta + \frac{\alpha^2}{2}\right] \left(\frac{\ln \omega_t - \ln \Omega_t^{(i)}}{s_t^2}\right)^2$$

$$J_{\ln \omega_r}(\{r_i\}) = \sum_i \left[\frac{r_i}{\alpha r_i + \beta}\right]^2 \left[\alpha r_i + \beta + \frac{\alpha^2}{2}\right] \left(\frac{\ln \omega_r - \ln \Omega_r^{(i)}}{s_r^2}\right)^2$$

11.1 Fisher information for direction of coherent motion

As a second example, consider a pattern of coherently moving wavelets k. The pattern has a speed v_0 and direction θ_0 and the temporal frequency of an individual wavelet $\omega_t^{(k)}$ will depend on its spatial frequency $\omega_r^{(k)}$ and its direction $\theta^{(k)}$ according to the cosine rule:

$$\omega_t^{(k)} = \omega_r^{(k)} \, v^{(k)} \qquad\qquad v^{(k)} = v_0 \, \cos\left(\theta^{(k)} - \theta_0\right)$$

$$\ln \omega_t^{(k)} = \ln \omega_r^{(k)} + \ln v_0 + \ln \cos\left(\theta^{(k)} - \theta_0\right) \qquad\qquad \theta^{(k)} = \theta_0 + \cos^{-1} \frac{\omega_t^{(k)}}{v_0 \, \omega_r^{(k)}}$$

$$\frac{\delta \ln \omega_t^{(k)}}{\delta \theta_0} = \tan\left(\theta^{(k)} - \theta_0\right) \qquad\qquad \frac{\delta \theta^{(k)}}{\delta \theta_0} = 1$$

We assume that we know the Fisher information $J_\theta^{(k)}$ for $\theta^{(k)}$ and $J_{\ln \omega_t}^{(k)}$ for $\ln \omega_t^{(k)}$ and would like to compute the Fisher information for the pattern parameter θ_0:

$$J_{\theta_0}^{(k)} = \left[\frac{\delta \theta^{(k)}}{\delta \theta_0}\right]^2 J_\theta^{(k)} + \left[\frac{\delta \ln \omega_t^{(k)}}{\delta \theta_0}\right]^2 J_{\ln \omega_t}^{(k)} = J_\theta^{(k)} + \tan^2\left(\theta^{(k)} - \theta_0\right) \, J_{\ln \omega_t}^{(k)}$$

Accordingly, the information about global direction derives from two components. One component is constant and equals the information about local direction. The other component depends on the maximal temporal frequency, $\omega_t^{max} = \omega_r^{(k)} \, v_0$, and on the sine of the angle between local and global direction, $\tan\left(\theta^{(k)} - \theta_0\right)$. The second component increases sharply with said angle and becomes singular for wavelets oriented orthogonally to the global direction.

For wavelets tuned to linear temporal frequency ω_t, the analogous formula is

$$J_{\theta_0}^{(k)} = \left[\frac{\delta \theta^{(k)}}{\delta \theta_0}\right]^2 J_\theta^{(k)} + \left[\frac{\delta \omega_t^{(k)}}{\delta \theta_0}\right]^2 J_{\omega_t}^{(k)} = J_\theta^{(k)} + \left[\omega_r^{(k)} \, v_0 \, \sin\left(\theta^{(k)} - \theta_0\right)\right]^2 \, J_{\omega_t}^{(k)}$$