THEORETICAL NEUROSCIENCE I

Lecture 14: Bayes' rule and maximum likelihood

Prof. Jochen Braun

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

Content

1. Motivation

2. Conditional probability and Bayes' rule

3. Example I: Scottish high school

4. Example II: Colored dice

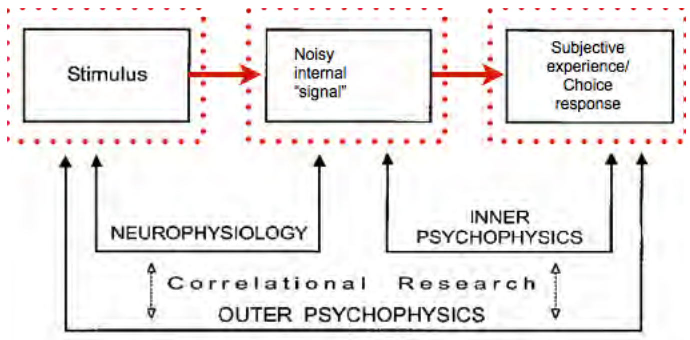
5. Joint decoding of two different neurons

6. Decoding a heterogeneous population

1 Meaning of neural activity

Neural activity $\longrightarrow$ Subjective experience $\longrightarrow$ Choice response

Physical cause $\longleftarrow$ Sensory signal $\longleftarrow$ **Neural activity**



Encoding and decoding

Encoding: sensory stimulus $\longrightarrow$ *variable neural activity*

Sensory stimuli evoke variable activity. Average evoked activity depends systematically on stimulus properties. We desire to know *probability* of different levels of activity, given different sensory stimuli.

Decoding: variable neural activity $\longrightarrow$ *stimulus estimate*

Variable neural activity determines choice response, which represents a (coarse) stimulus alternative. More generally, neural activity lets us estimate stimulus properties. The reliability of such estimates depends on encoding. We desire to know the *likelihood* of different stimulus properties, given the neural activity evoked by different stimuli.

2 Bayes' Rule

We consider two random variables and ask, what does the value of one tells us about the other (statistically speaking). To this end, we will employ "Bayes' Rule" or "Bayes' Law".

For example, the two random variables could be a stimulus value and the activity of a neuron, which we happen to be observing simultaneously. In this case, we would like to know what the neuron's activity at a given time tells us about the stimulus value at the same moment.

Figure 1: Thomas Bayes, 1701-1761. [1]

Joint probability $P(r, s)$
(dt.: gemeinsame od. Verbundwahrscheinlichkeit)

The *joint probability* is the probability that *two* random variables simultaneously assume particular values. It depends on the values of both random variables and is thus a two-dimensional function.

r *and* **s**

Marginal probabilities $P(r)$, $P(s)$ (dt.: a priori Wahrscheinlichkeit)

The *marginal probability* is the probability that *one* random variable assumes a particular value, independent of the value of the other. It is a one-dimensional function and can be obtained from the *joint probability* by summation (discrete case) or integration (continuous case)

r *alone, irrespective of* **s**

s *alone, irrespective of* **r**

Conditional probabilities $P(r|s)$, $P(s|r)$ (dt. bedingte od. a posteriori Wahrscheinlichkeit)

The *conditional probability* is defined for subset of outcomes (the *given condition*), in which the second variable takes a particular value. It is the probability that the first variable takes different values. It is a two-dimensional function.

r *given* **s**

s *given* **r**

General relations

The joint probability $P[r, s]$ contains all other probabilities. To compute marginal probabilities $P[r]$ and $P[s]$, we sum $P[r, s]$ over either r or s:

$$P[r] = \sum_s P[r, s] \qquad P[s] = \sum_r P[r, s]$$

To obtain the conditional probabilities $P[r|s]$ and $P[s|r]$, we divide $P[r,s]$ by either $P[r]$ or $P[s]$:

$$P[r|s] = \frac{P[r,s]}{P[s]} \qquad\qquad P[s|r] = \frac{P[r,s]}{P[r]}$$

Bayes' Rule

With these definisitions we obtain

$$P[r|s]\,P[s] = P[r,s] \qquad\qquad P[r,s] = P[s|r]\,P[r]$$

and, further

$$P[r|s]\,P[s] = P[s|r]\,P[r] \qquad \Leftrightarrow$$

$$P[r|s] = \frac{P[s|r]\,P[r]}{P[s]} \qquad \Leftrightarrow \qquad P[s|r] = \frac{P[r|s]\,P[s]}{P[r]}$$

These relations are called **Bayes' Rule**.

3 Example I: Scottish high-school

In a Scottish high-school, the *joint probabilities* of boys and girls wearing trousers and skirts are

Figure 2: Scottish high-school. [2]

	Boys	Girls	Total
Skirts	0.1	0.2	
Trousers	0.5	0.2	
Total			1.0

The *marginal probabilities* can be obtained by summing along rows or columns:

	Boys	Girls	Total
Skirts	0.1	0.2	0.3
Trousers	0.5	0.2	0.7
Total	0.6	0.4	1.0

Conditional probabilities are obtained as a ratio of joint and marginals:

$$P(\text{girl}|\text{skirt}) = \frac{P(\text{girl}, \text{skirt})}{P(\text{skirt})} = \frac{0.2}{0.3} = 0.66$$

$$P(\text{skirt}|\text{girl}) = \frac{P(\text{girl}, \text{skirt})}{P(\text{girl})} = \frac{0.2}{0.4} = 0.50$$

Conditional probability of gender, given dress:

	Boys	Girls	Total
Skirts	1/3	2/3	1
Trousers	5/7	2/7	1

	Boys	Girls	Total
Skirts	0.1	0.2	0.3
Trousers	0.5	0.2	0.7
Total	0.6	0.4	1.0

Conditional probability of dress, given gender:

	Boys	Girls	
Skirts	1/6	1/2	
Trousers	5/6	1/2	
Total	1	1	

	Boys	Girls	Total
Skirts	0.1	0.2	0.3
Trousers	0.5	0.2	0.7
Total	0.6	0.4	1.0

4 Example II: colored dicel

Consider *four red* and *five white* dice. The total number of eyes on the red dice lies between 4 and 24 and is a discretely distributed random variable. The number of eyes on the white dice lies between 5 and 30 and is another such variable.

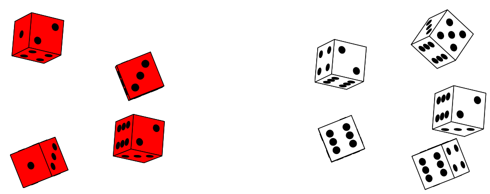


Figure 3: Color dice.

Independent random variables

Red and white dice are **independent!** Thus, the joint probability of outcomes is simply the product of the individual probabilities:

$$P(n_{red}, n_{white}) = P(n_{red}) \cdot P(n_{white})$$

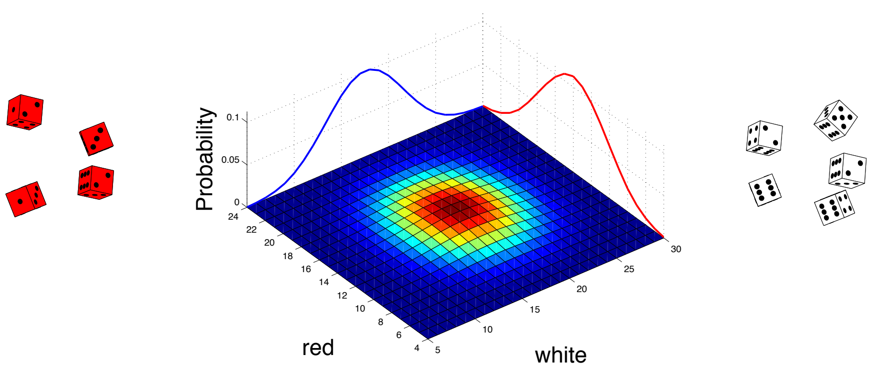


Figure 4: Joint probabilities of independent variables.

The marginal probabilities are obtained by summation:

$$P(n_{red}) = \sum_{n_{white}} P(n_{red}, n_{white}) \qquad\qquad P(n_{white}) = \sum_{n_{red}} P(n_{red}, n_{white})$$

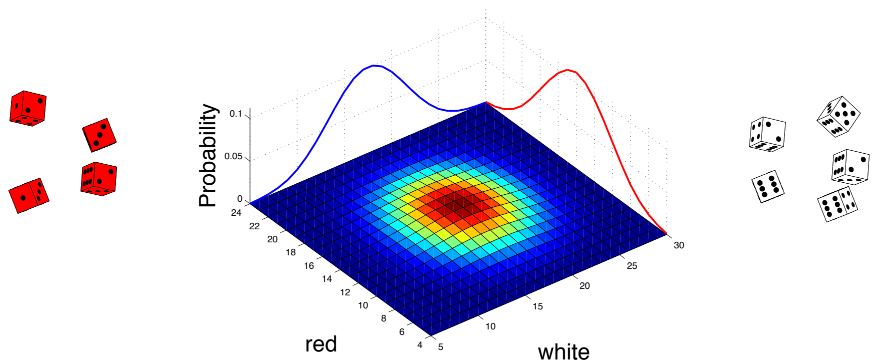


Figure 5: Marginal probabilities.

The conditional probability of n_{red} eyes, given that the *white* dice show n_{white} eyes, is

$$P(n_{red}|n_{white}) = \frac{P(n_{red}, n_{white})}{P(n_{white})}$$

Because of **independence**, conditional and marginal probabilities are the same:

$$P(n_{red}|n_{white}) = \frac{P(n_{red})\, P(n_{white})}{P(n_{white})} = P(n_{red})$$

$$P(n_{red}|n_{white}) = \frac{P(n_{red}, n_{white})}{P(n_{white})} = \frac{P(n_{red})\, P(n_{white})}{P(n_{white})} = P(n_{red})$$

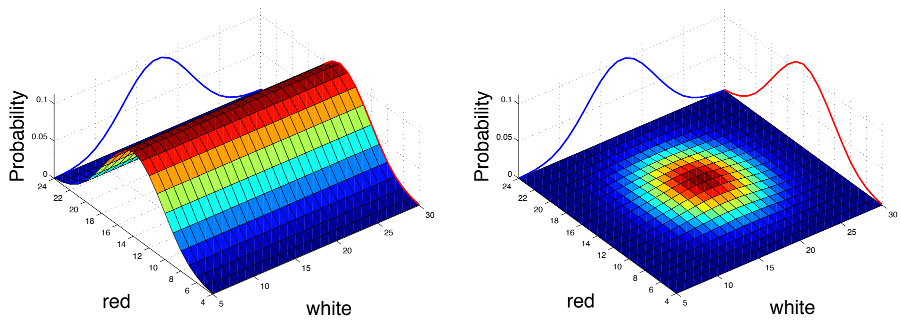


Figure 6: Probabilities of red given white.

$$P(n_{white}|n_{red}) = \frac{P(n_{red}, n_{white})}{P(n_{red})} = \frac{P(n_{red})\,P(n_{white})}{P(n_{red})} = P(n_{white})$$

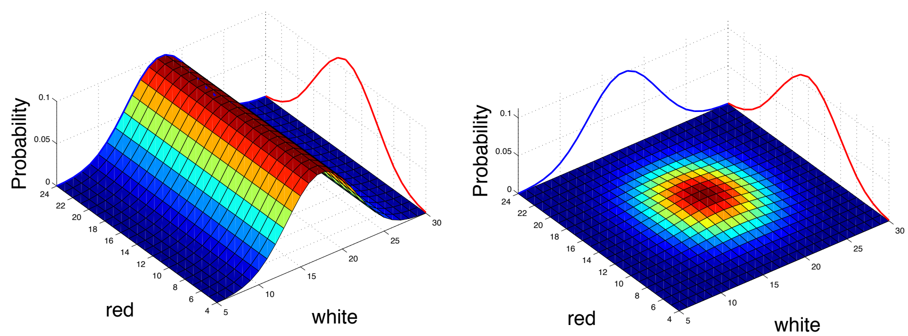


Figure 7: Probabilities of white given red.

Dependent random variables

Now consider *one red*, *three blue* and *two white* dice. The total number of eyes on the *red and blue* dice lies between 4 and 24 and is a discretely distributed random variable. The number of eyes on the *blue and white* dice lies between 5 and 30 and is another such variable.

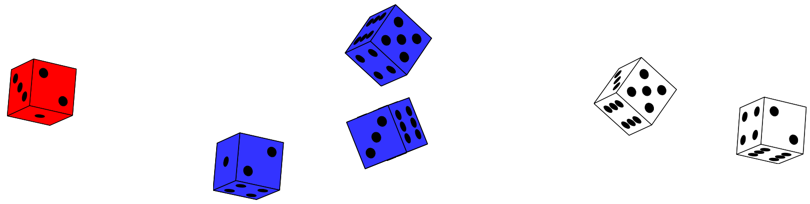


Figure 8: Three colors dice.

As the *blue* dice belong to both groups, the joint probability of outcomes is no longer simply the product of the individual probabilities:

$$P(n_{red\&blue}, n_{blue\&white}) \neq P(n_{red\&blue}) \cdot P(n_{blue\&white})$$

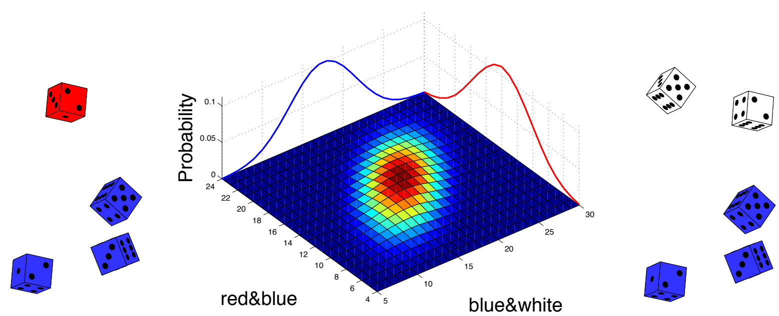


Figure 9: Probabilities of the three colors dice.

The marginal probabilities are obtained by summation:

$$P(n_{red\&blue}) = \sum_{n_{blue\&white}} P(n_{red\&blue}, n_{blue\&white})$$

$$P(n_{blue\&white}) = \sum_{n_{red\&blue}} P(n_{red\&blue}, n_{blue\&white})$$

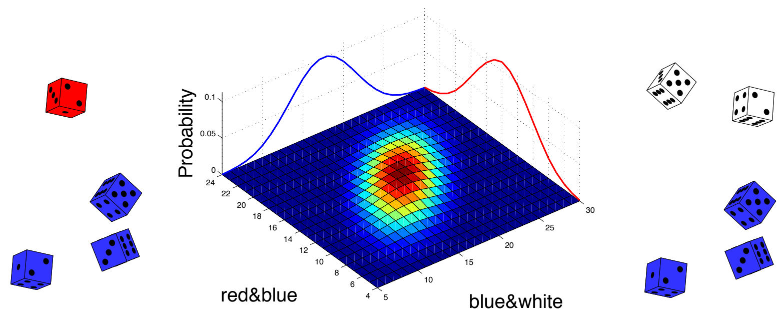


Figure 10: Marginal probabilities.

$$P(n_{red\&blue}|n_{blue\&white}) = \frac{P(n_{red\&blue}, n_{blue\&white})}{P(n_{blue\&white})}$$

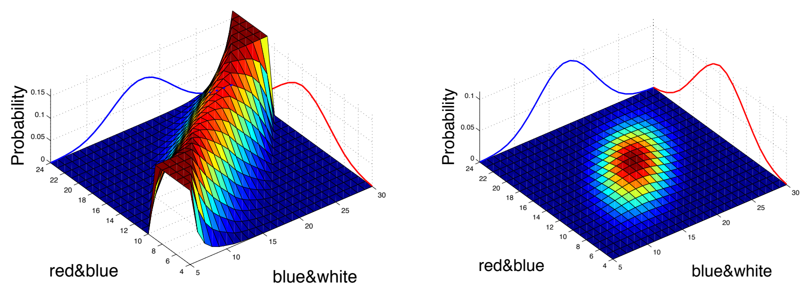
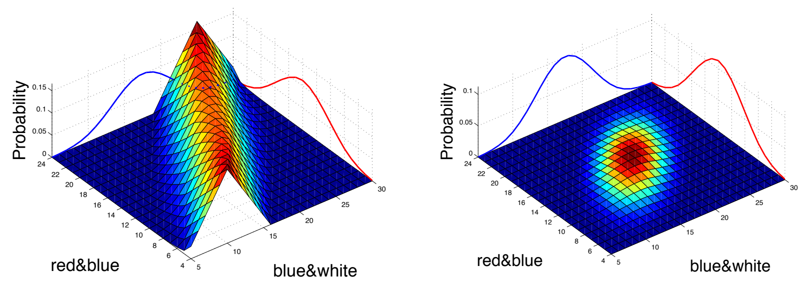


Figure 11: Probabilities of red and blue given blue and white.

$$P(n_{blue\&white}|n_{red\&blue}) = \frac{P(n_{red\&blue}, n_{blue\&white})}{P(n_{red\&blue})}$$



Summary Bayes' Rule

Joint

$$P[r, s]$$

Marginal

$$P[r] = \sum_s P[r, s] \qquad P[s] = \sum_r P[r, s]$$

Conditional

$$P[r|s] = \frac{P[r, s]}{P[s]} \qquad P[s|r] = \frac{P[r, s]}{P[r]}$$

$$P[r|s]\, P[s] = P[r, s] \qquad P[s|r]\, P[r] = P[r, s]$$

Bayes' rule

$$P[r|s]\, P[s] = P[s|r]\, P[r]$$

5 Joint decoding of two neurons

We now apply these conditional probability and Bayes' rule to decoding neural responses. In general, we have a heterogeneous population, in which each neuron preferes a different stimulus attribute (*e.g.*, a different direction of motion).

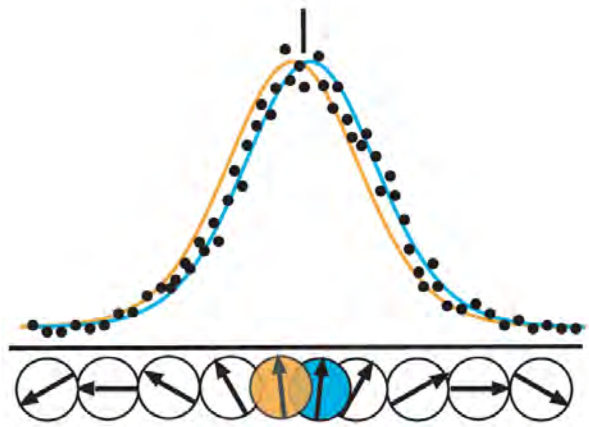


Figure 12: Different directions of motion.

Combine likelihoods, rather than responses!

We cannot simply combine responses across neurons (as the total response would be uninformative about the stimulus). Instead, we need to infer from each individual response a stimulus *likelihood*, which we can then combine across the population.

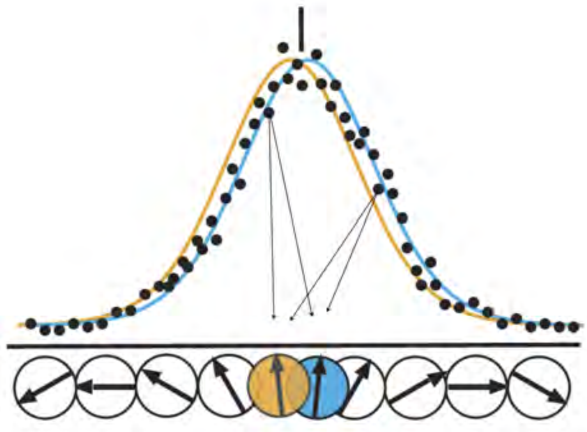


Figure 13: Different directions of motion.

Conditional response probability

As an example, we consider neurons i tuned for directions of visual motion. We approximate the **tuning curve** by a Gaussian function

$$r_i = f_i(\theta) = r_{max} \exp\left[-(\theta - \theta_i)^2/2\kappa^2\right]$$

where θ_i is the preferred direction of neuron i and κ determines the tuning width.

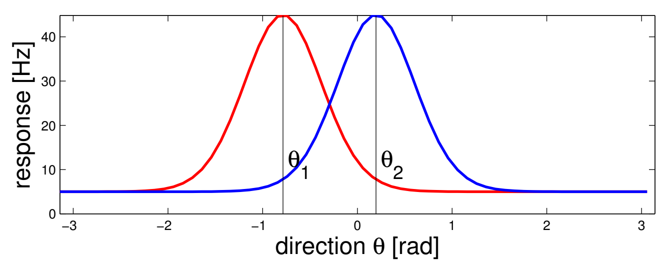


Figure 14: Aproximation of tuning curve by a Gaussian function.

Assuming independence, we approximate the **response variability** with a Poisson distribution

$$p(x|r_i) = \frac{(r_i)^x}{x!} e^{-r_i}$$

where x is the actual and r_i is the average response:

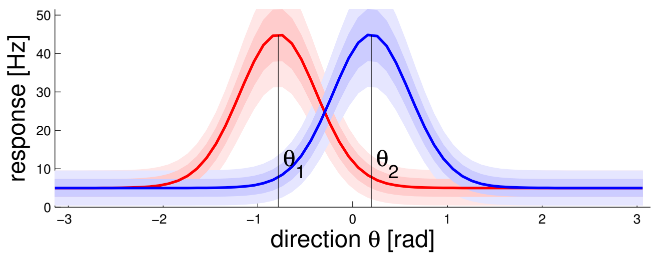


Figure 15: Aproximation of the response variability with a Poisson distribution.

Note: while the Poisson assumption is realistic, the independence assumption is not.

Combining tuning function and response variability, we obtain the **conditional response probability:**

$$p(x|\theta) = \frac{(r_i)^x}{x!}\, e^{-r_i} = \frac{[f_i(\theta)]^x}{x!} e^{-f_i(\theta)}$$

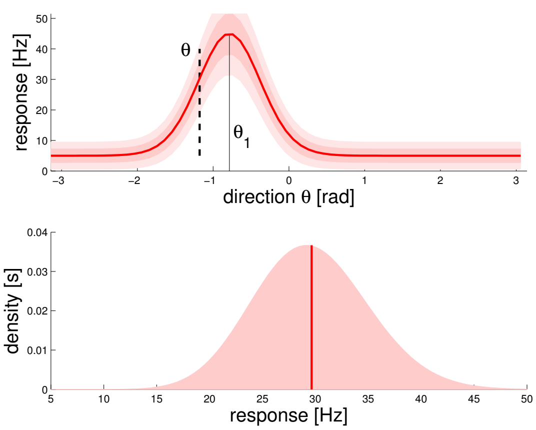


Figure 16: Conditional response probability.

Stimulus likelihood

If all possible stimuli θ are equally probable, the conditional response probability can be re-interpreted as a **stimulus likelihood**

$$L(\theta|x_1) = p(x_1|\theta) = \frac{[f_1(\theta)]^{x_1}}{x_1!}\, e^{-f_1(\theta)}$$

namely, the *likelihood* that an observed response x was caused by stimulus θ.

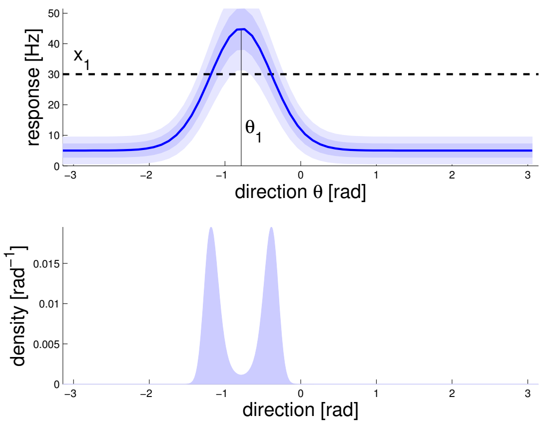


Figure 17: Stimulus likelihood.

Multiple stimulus likelihoods

From multiple responses (of different neurons), we obtain multiple likelihoods

$$L(\theta|x_1) \qquad\qquad L(\theta|x_2)$$

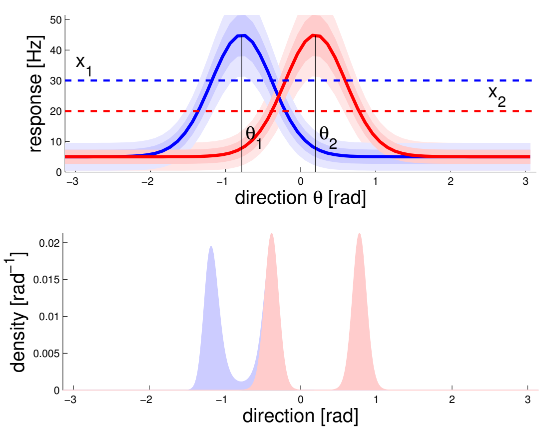


Figure 18: Multiple stimulus likelihoods

Combining likelihoods

Assuming that the response variability of different neurons is independent, we can compute the **joint likelihood** that stimulus θ caused both observed responses x_1 and x_2 as the product of the individual likelihoods

$$L(\theta|\{x_1, x_2\}) = L(\theta|x_1)\, L(\theta|x_2)$$

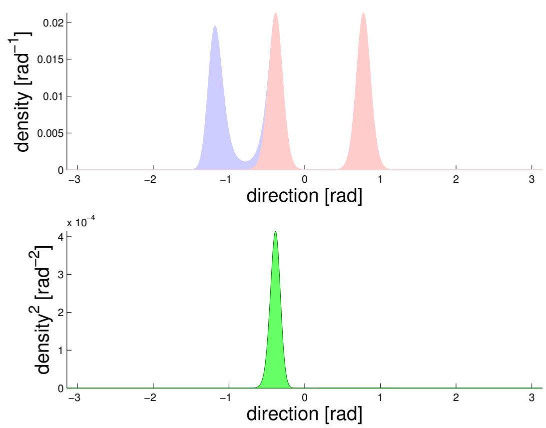


Figure 19: Combining likelihoods.

Summary joint decoding of two neurons

- **The combined response of a heterogeneous population is uninformative about the stimulus.**

- **To proceed, we rely on prior information about each neuron i, namely, its conditional response probability $P(x_i|\theta)$.**

- **This equals the likelihood $L(\theta|x_i)$ of an observed response x_i being caused by stimulus θ, provided all θ are equally probable.**

- **The joint likelihood $L(\theta|x_1, x_2)$ of an observed response x_1, x_2 being caused by stimulus θ, is obtained by multiplication**

$$L(\theta|x_1, x_2) = L(\theta|x_1)\, L(\theta|x_2)$$

6 Generalization to many neurons (advanced)

In general, the likelihood of stimulus θ, given an observed population response $\{x_1, x_2, \ldots x_n\}$, is

$$L(\theta|\{x_1, x_2, \ldots x_n\}) = \prod_i L(\theta|x_i)$$

if response variability is independent.

To find the most likely stimulus θ, we seek the maximum of this function or, equivalently, the maximum of its logarithm (**'log likelihood'**) :

$$\log L(\theta|\{x_1, x_2, \ldots x_n\}) = \log \prod_i L(\theta|x_i) = \sum_i \log L(\theta|x_i)$$

The logarithm is convenient because sums are easier to compute than products.

Poisson variability

Assuming Poisson-distributed variability, we have

$$L(\theta|x_i) = \frac{f_i(\theta)^{x_i}}{x_i!} e^{-f_i(\theta)}$$

$$\log L(\theta|x_i) = x_i \log f_i(\theta) - \log x_i! - f_i(\theta)$$

$$\sum_i \log L(\theta|x_i) = \sum_i x_i \log f_i(\theta) - \sum_i \log x_i! - \sum_i f_i(\theta)$$

which simplifies to

$$\log L(\theta|\{x_1, x_2, \ldots x_n\}) = \sum_i \log L(\theta|x_i) \quad \propto \quad \sum_i x_i \log f_i(\theta)$$

as the other terms are independent of θ.

Aside: independence of θ?

The second term

$$\sum_i \log x_i!$$

depends only on the observed response $\{x_1, \ldots, x_n\}$, not on θ.

The third term

$$\sum_i f_i(\theta)$$

also does not depend on θ, provided the neural population covers all possible stimuli ('uniform coverage').

Aside: single population response

Mean response: dashed red line

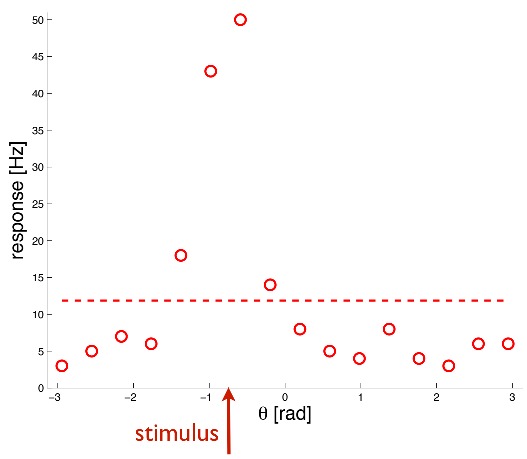


Figure 20: Single population response.

Aside: two population responses

Mean responses (dashed red and blue lines) do not depend systematically on stimulus θ!

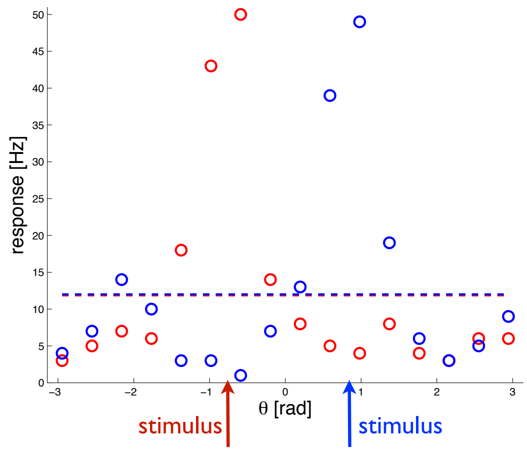


Figure 21: Two population responses.

Maximum likelihood stimulus

The most likely stimulus $\theta = \theta_{est}$ to have caused an observed response $\{x_1, \ldots, x_n\}$ is the value that maximizes

$$\log L(\theta | \{x_1, x_2, \ldots x_n\}) = \sum_i x_i \, \log f_i(\theta)$$

or, in other words, the response-weighted average of the log-tuning $\log f_i(\theta)$.

Compare **observed response** to logarithmic mean of **hypothetical response** to a hypothetical stimulus θ!

The better the match between observed response and hypothetical response, the more likely the hypothetical stimulus!

Gaussian tuning

Assuming Gaussian tuning we find

$$f_i(\theta) = r_{max} \, e^{-\frac{(\theta - \theta_i)^2}{2\kappa^2}}$$

$$\log f_i(\theta) = -\frac{(\theta - \theta_i)^2}{2\kappa^2} + \log r_{max}$$

$$\frac{\partial}{\partial\theta} \log f_i(\theta) = -\frac{\theta - \theta_i}{\kappa^2}$$

Maximum likelihood estimate

To determine this, we demand

$$\frac{\partial}{\partial\theta} \log L(\theta|\{x_1, x_2, \ldots x_n\}) \stackrel{!}{=} 0$$

$$\frac{\partial}{\partial\theta} \sum_i x_i \, \log f_i(\theta) = \sum_i x_i \frac{\partial}{\partial\theta} \log f_i(\theta) = \sum_i x_i \left(-\frac{\theta - \theta_i}{\kappa^2} \right) \stackrel{!}{=} 0$$

$$-\frac{1}{\kappa^2} \sum_i x_i \, (\theta_{ml} - \theta_i) = 0 \qquad \Leftrightarrow \qquad \theta_{ml} \sum_i x_i - \sum_i x_i \, \theta_i = 0$$

$$\theta_{ml} = \frac{\sum_i x_i \, \theta_i}{\sum_i x_i}$$

θ_{ml} is the maximum likelihood estimate of the stimulus that caused the observed population response. For Poisson variability and Gaussian tuning, it works out as the response-weighted average of the preferred stimuli θ_i.

Summary generalization to many neurons

- To decode a heterogeneous population, we rely on the conditional response probability of each neuron i $P(x_i|\theta) \approx L(\theta|x_i)$.
- The joint likelihood $L(\theta|\{x_1, \ldots, x_n\})$ of an observed response $x_1, \ldots, x_n$ being caused by stimulus θ, is

$$L(\theta|x_1, \ldots, x_n) = L(\theta|x_1) \, L(\theta|x_2) \, \ldots \, L(\theta|x_n)$$

- The 'maximum likelihood estimate' θ_{est} is obtained by maximizing

$$\prod_i L_i(\theta|x_i) \quad \text{or} \quad \sum_i \log L_i(\theta|x_i)$$

- The ML estimate is the response-weighted average

$$\theta_{est} = \frac{\Sigma_i \; x_i \, \theta_i}{\Sigma_i \; x_i}$$

(neurons with Gaussian tuning and Poisson variability).

7 Bibliography

1. Thomas Bayes, Anonymous. Ref: http://mnstats.morris.umn.edu/introstat/history/w98/Bayes.html
2. Glasgow School Photographs. Fine Art America. Ref: https://fineartamerica.com/featured/old-ibrox-public-school-govan-parish-school-board-glasgow-scotland-uk-joe-fox.html