Stochastic models in neurocognition

Patricia.Reynaud-Bouret@univ-cotedazur.fr

Etienne.Tanre@inria.fr

Josue.Tchouanti-Fotso@univ-cotedazur.fr

1st 3 courses : Patricia independent models

2 courses : Etienne Markov chains and Poisson process

2 courses : Patricia point processes and their statistics

3 —— Etienne PDMP, Brownian motion, mean-field

Models with independance and their statistical tools

1) Some Examples of models for neurocognition with independance

→ independance helps us to do a lot of ≠ statistical analysis

→ when facing a model / data to model, you have to identify where the independance is

→ individuals : cognitive experiment and different participant.

⚠ not true if the participants are interacting

→ trials : when you ask the participant to repeat the experiment

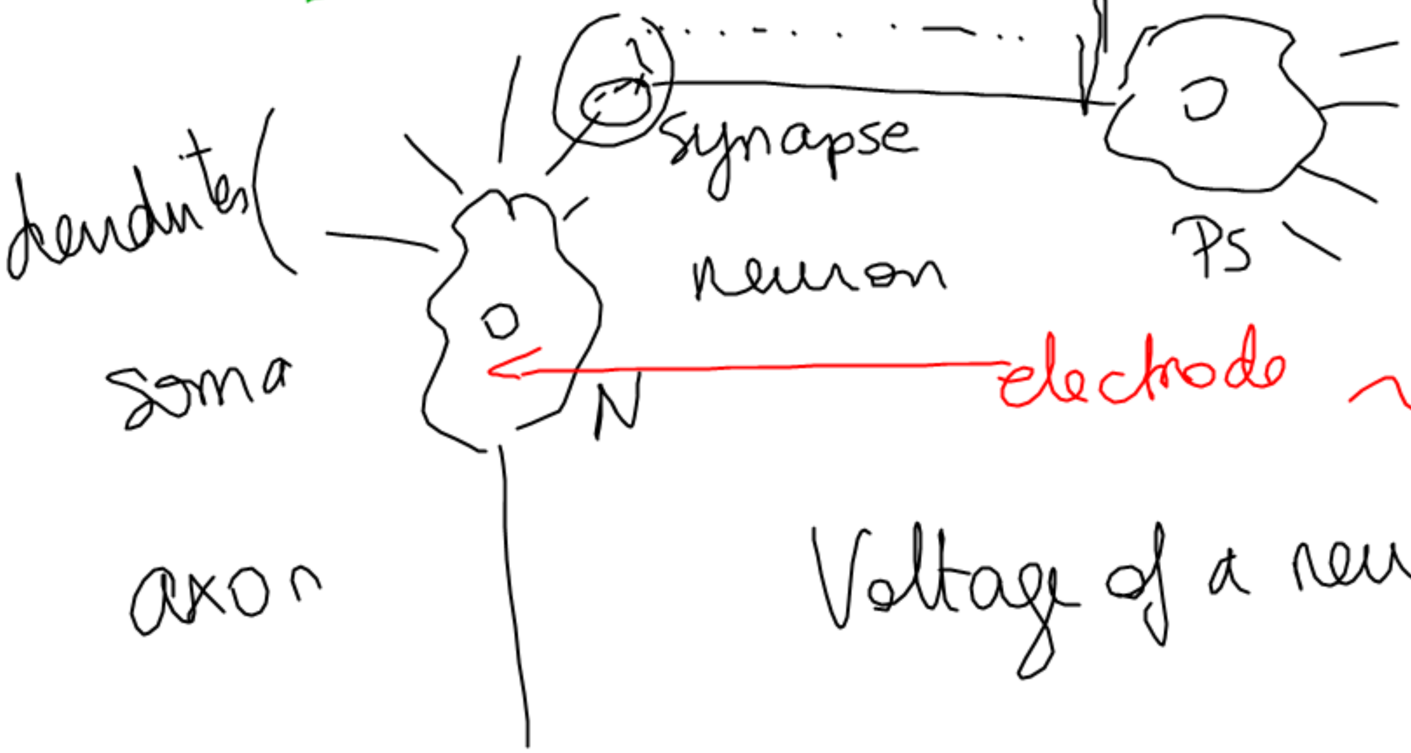
⚠ Be aware that people can get tired → not iid ...

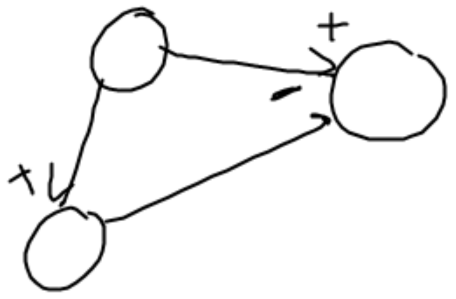
→ Another important feature is

parametric models vs non parametric models

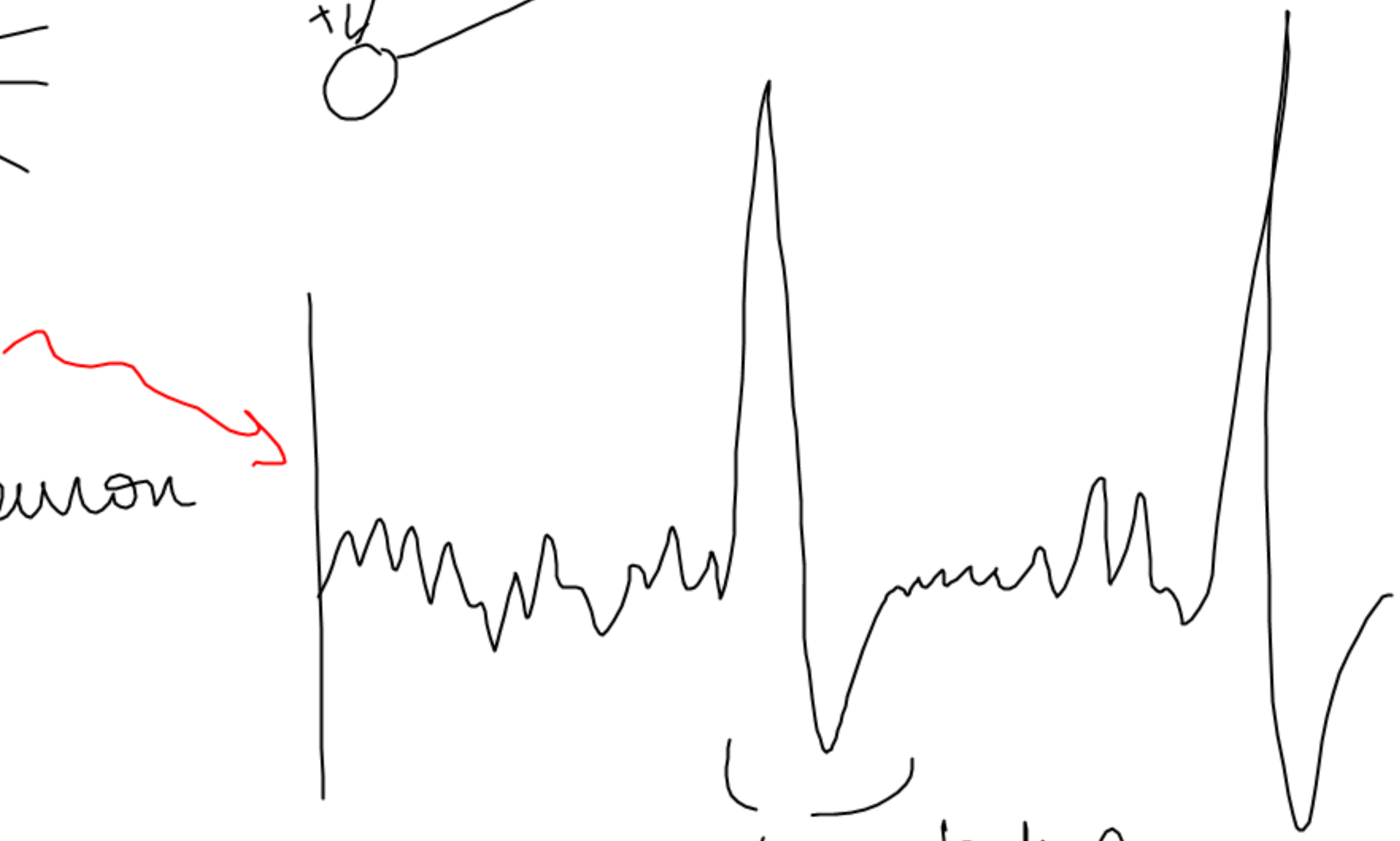
parametric models	non parametric models
The distribution of the data is parametrized by a finite number of parameters	The distribution depends on more than a finite set of parameters for instance it depends on a function f.
eg $\mathcal{N}(m, \sigma^2)$ $\theta = (m, \sigma^2) \in \mathbb{R}^2$	eg $X_1 \dots X_n$ iid with cumulative distribution function $F(\ )$

1) Interspikes intervals



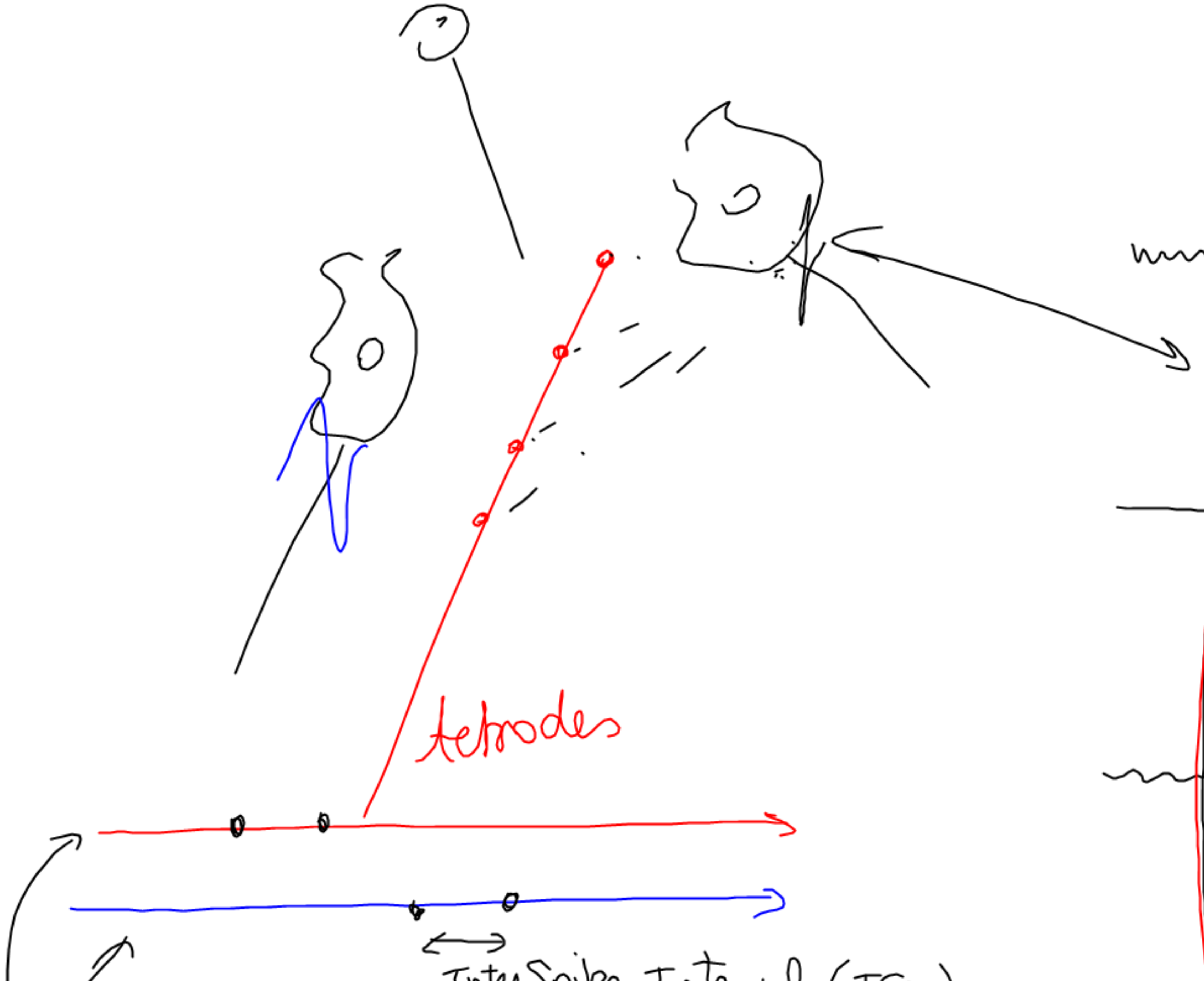


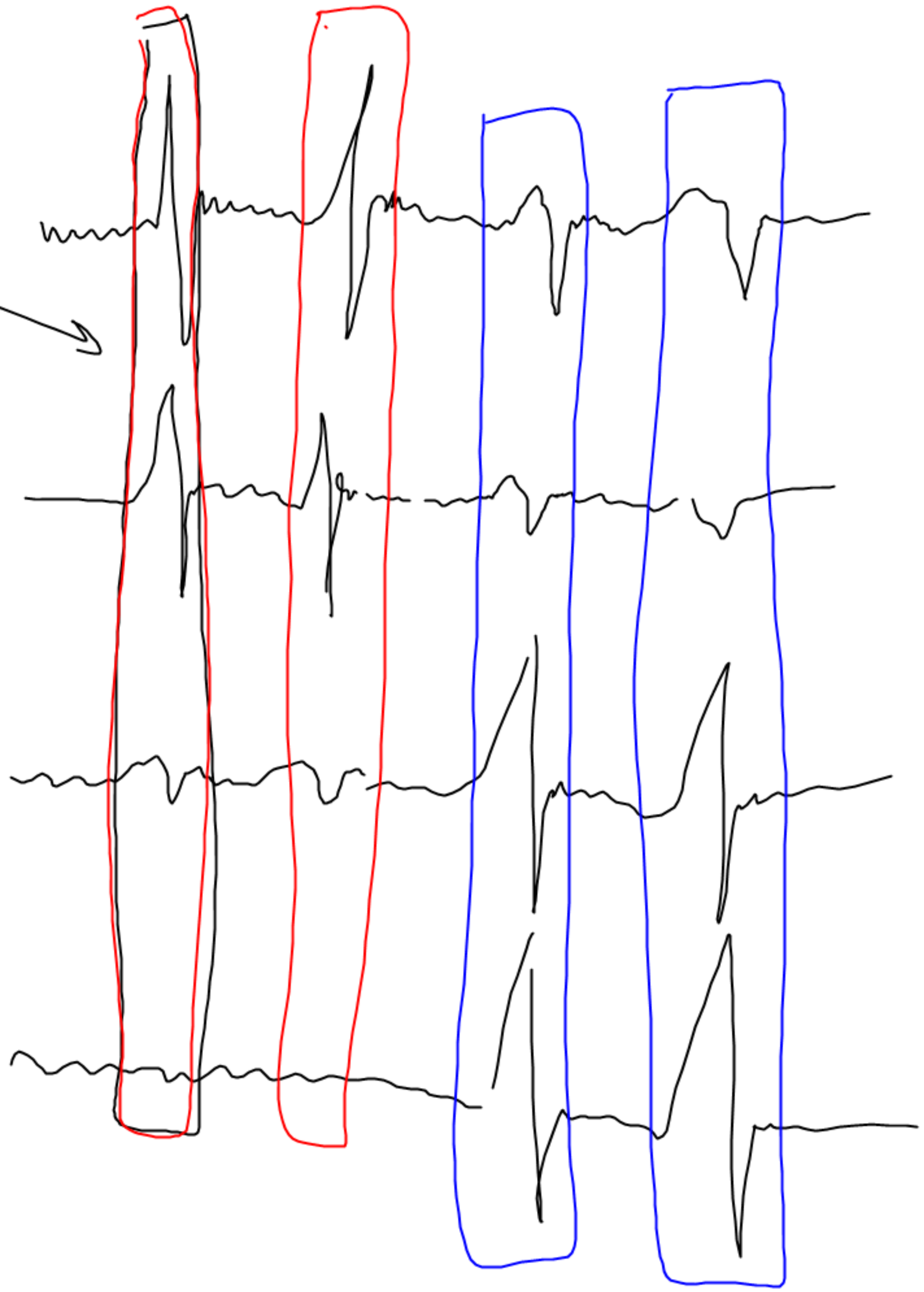
Voltage of a neuron



for a given neuron, the shape of the action potential is always the same.

presynaptic neuron is excitatory (inhibitory) a spike of it (PS) → higher (lower) Voltage for N → the higher the voltage is, the more likely N spikes





spike train

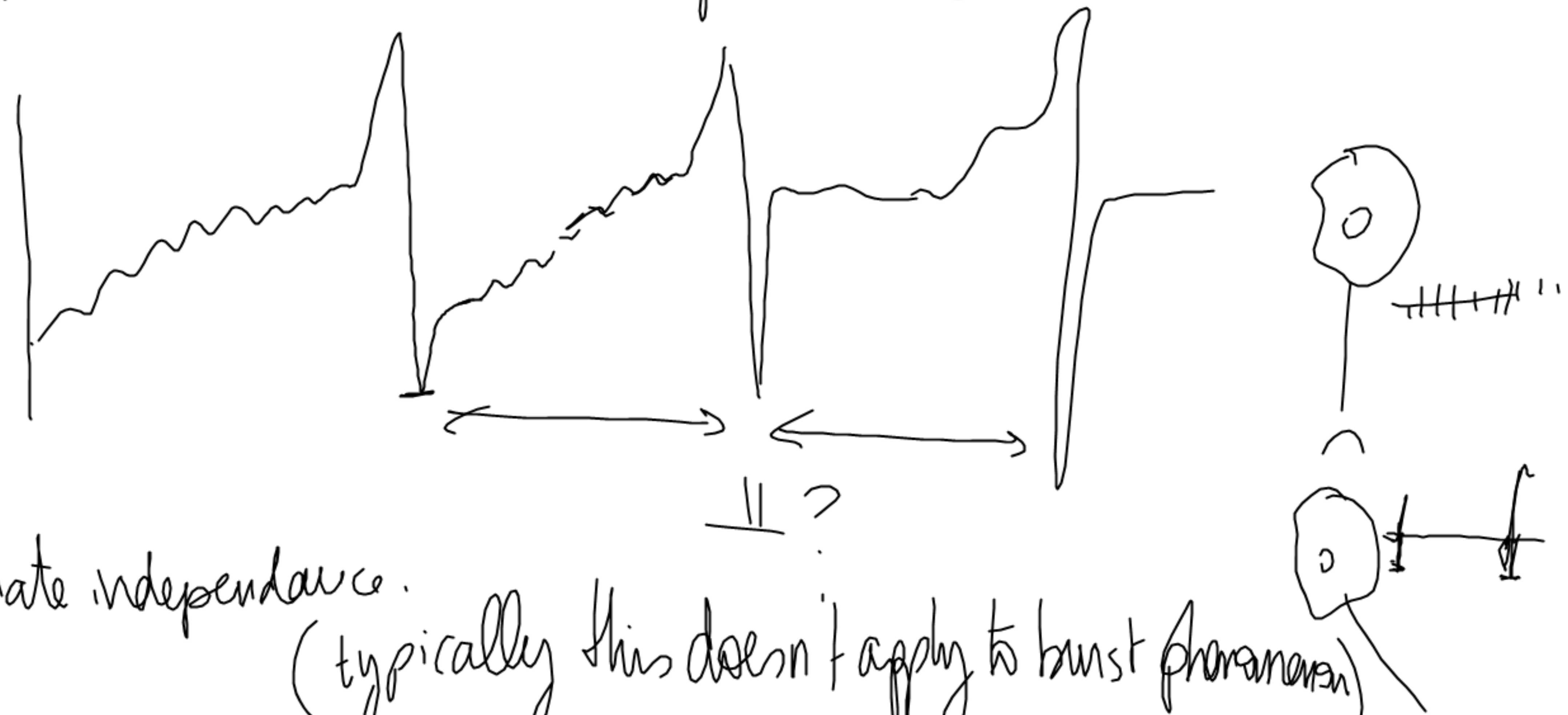
→ that's where the information ...

We can model the ISI's of a given neuron as i.i.d variables
(independent and identically distributed)

e.g. Bruno's course
(Leaky) Integrate-and-fire ⇝ a particular distribution for the ISI's

→ Independant

at least the voltage is next to the same value at each spike
⇝ this would legitimate independance.



(typically this doesn't apply to burst phenomenon)

→ identically distributed

→ true as long as the behavior of the animal does not change too much

→ small portion of time

<u>NB</u> : It means I do not model

- change in behaviors
- memory effects (capacitance, STDP...)
- interaction between neurons.

Now let us specify a bit more the models

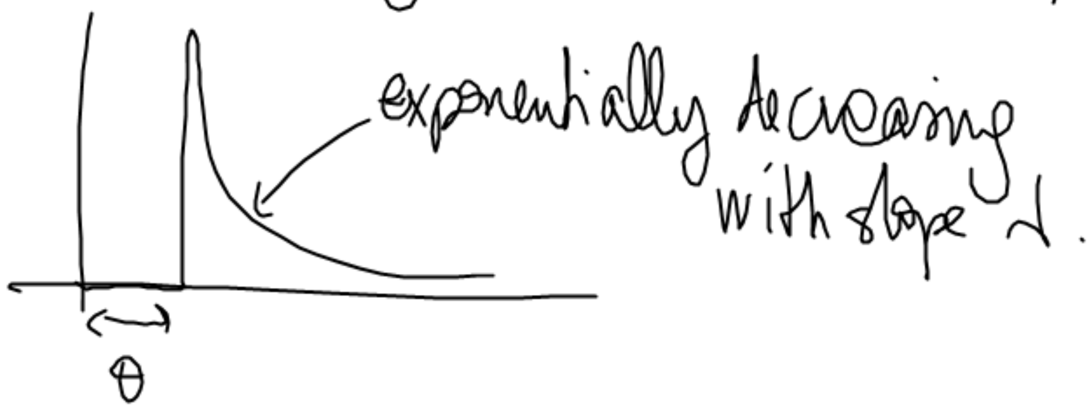
→ parametric assumption

- $X_1, \dots, X_n$ the ISI's are iid $\mathcal{E}(\lambda)$, λ is unknown (1 parameter)

($\rightsquigarrow$ Poisson processes, see Etienne's course)

- But this model does not take into account refractory period

→ shifted exponential density is given by $f(x) = \lambda e^{-(x-\theta)\lambda} \mathbb{1}_{x \geq \theta}$



<u>NB</u>: When λ is small, say 1 or 3 Hz, since the ref period $\simeq$ 2 ms to 5 ms we will not see this effect and the 1st model is good...

→ **non parametric assumption**

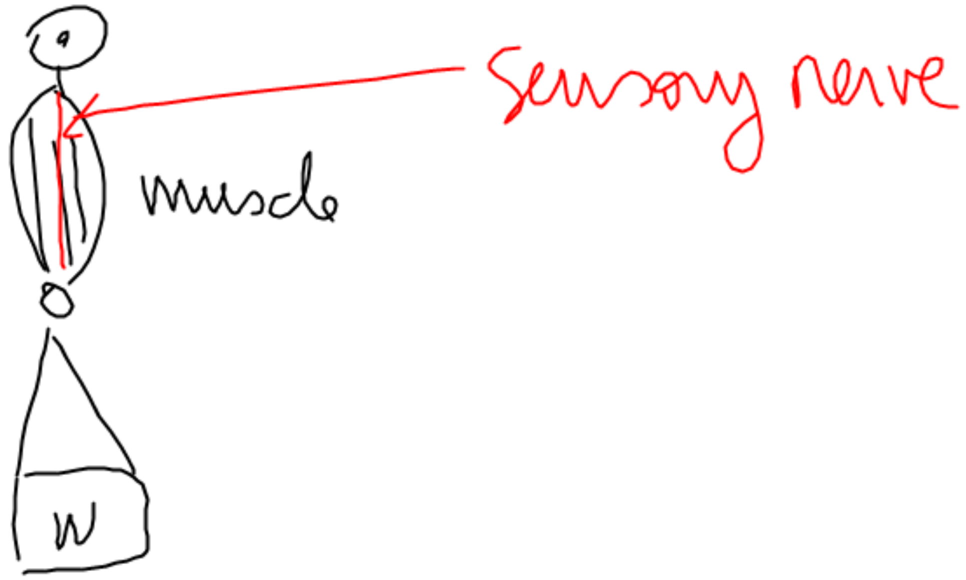
$X_1 \dots X_n$ are iid with density f, f unknown.

→ think of density( ) in R → an estimator $\hat{f}$ of f.

2) Neural Rate Coding

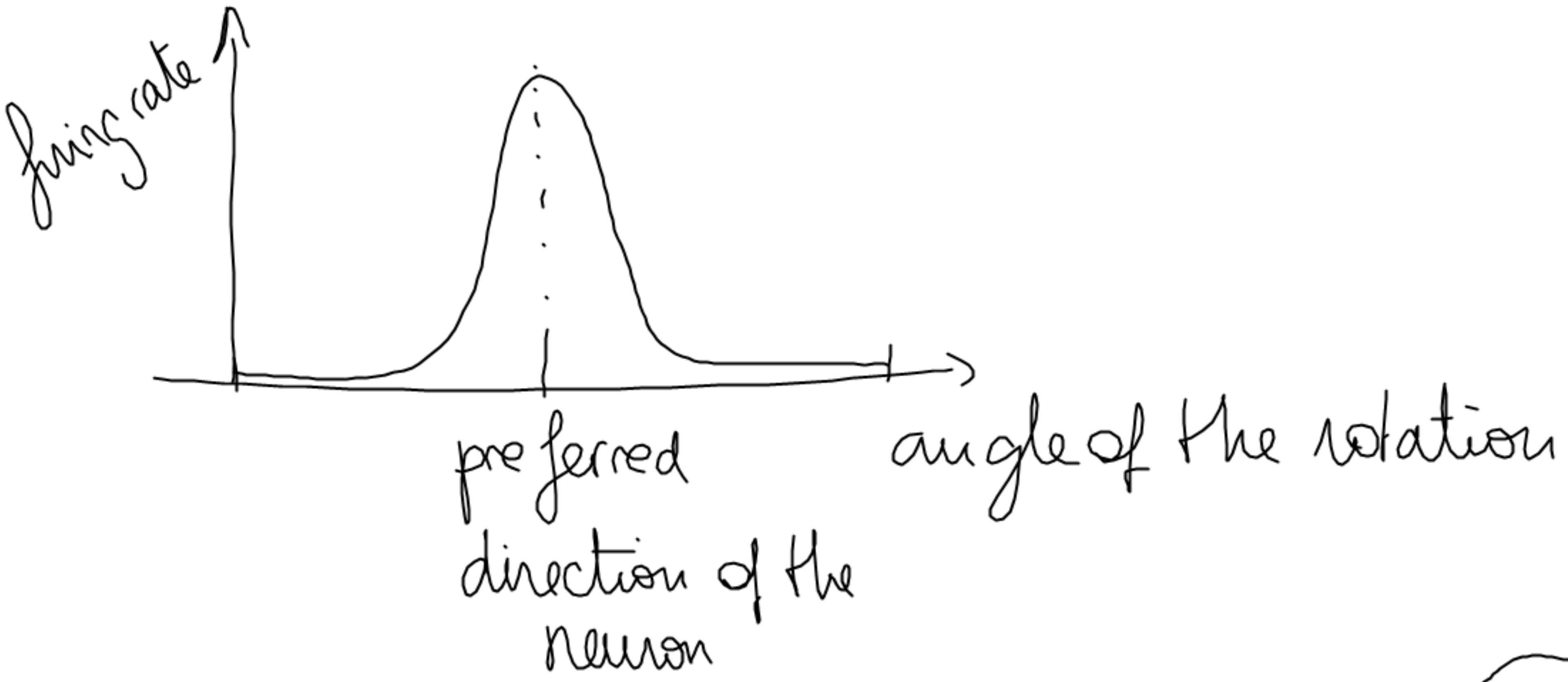
(the firing rate of a neuron is in average the number of spikes produced per second)

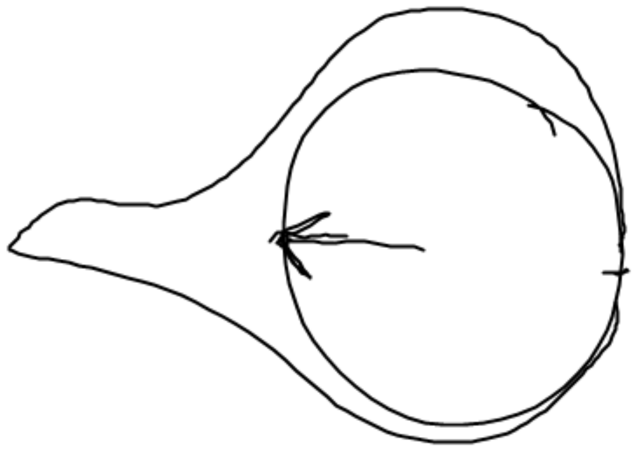
in 1926, Adrian and Zottermann



the firing rate of the sensory nerve increase with the weight

1986, Georgopoulos, Schwarz and Kettner





to make this into a model you say that
parametric

weight

$$\text{firing rate} = Y = \begin{cases} \underline{a + b\,W} + \sigma\varepsilon \\ \underline{a + b\cos\theta} + \sigma\varepsilon \end{cases}$$

a, b, σ unknown
$\varepsilon \sim \mathcal{N}(0,1)$

angle of the movement.
θ = angle between the movement and the preferred direct° of movement.

non parametric model

$$Y = \begin{cases} f(W) \\ f(\theta) \end{cases} + \sigma\varepsilon$$

with f unknown.

→ for one cell, I try several W or θ → the observation are independent.
→ for different cells → you "hope" that this is still independent.

What you should know,

- estimation (Law of Large Numbers)
- asymptotic confidence intervals (Central Limit Theorem)
- tests → parametric (lm())
  → nonparametric (ks.test(), wilcox.test(),
  shapiro.test()
  χ^2 tests)

II) Likelihood and contrast

1) Likelihood

Toy example

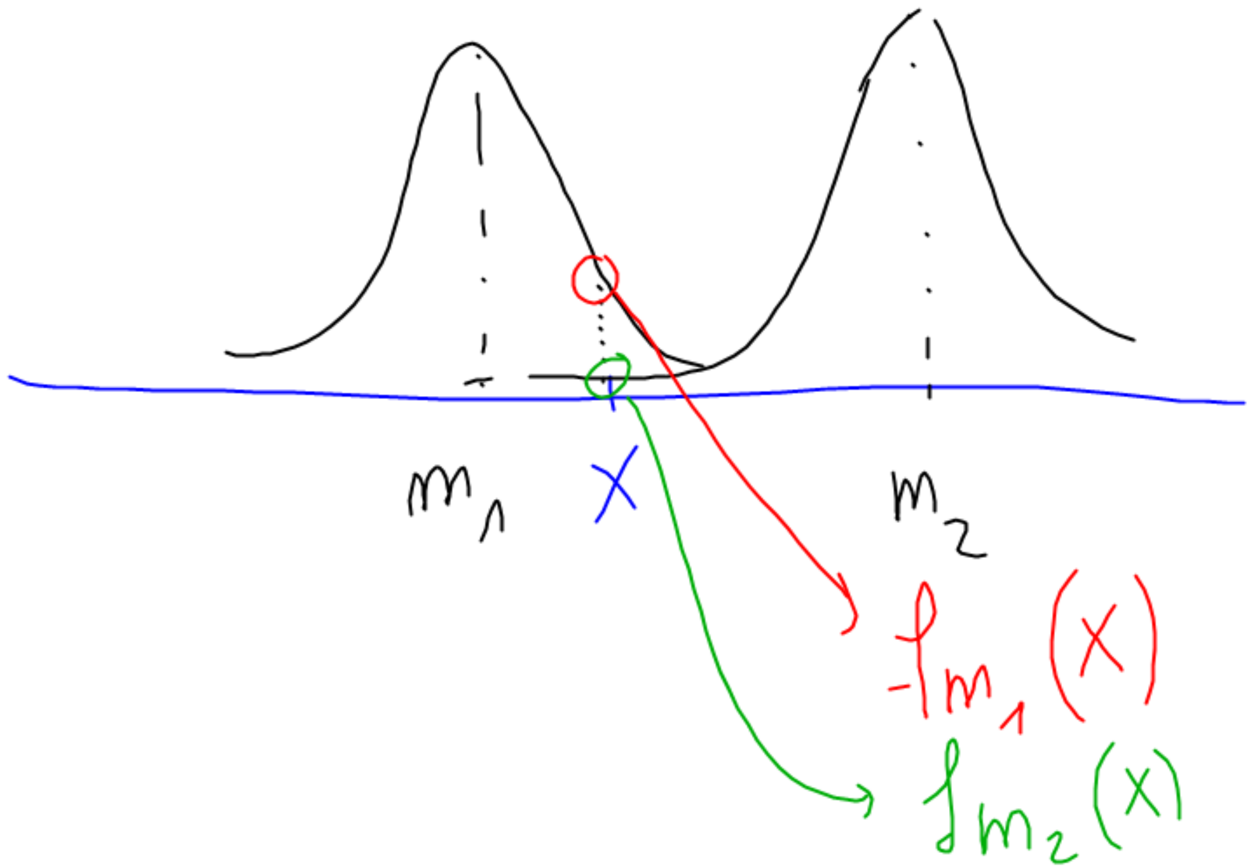
I observe $X \sim \mathcal{N}(m,1)$ and I have the choice between m_1 and m_2

$\Theta = \{m_1, m_2\}$

$\theta \mapsto f_\theta(X)$ has two values $f_{m_1}(X)$ and $f_{m_2}(X)$

the maximum is achieved in $\theta = m_1$

$\longrightarrow \hat{\theta}(MLE) = m_1$



Since $f_{m_1}(X) > f_{m_2}(X)$

$\longrightarrow$ I choose m_1

In general you have a parametric family $\{f_\theta\}$ parametrized by $\theta \in U \subset \mathbb{R}^d$

f_θ is either a density if the variable is continuous
or the probability distribution function if the variable is discrete.

you observe $X \sim f_{\theta_0}$ for θ_0 unknown

Likelihood: $\theta \longmapsto f_\theta(X)$

⚠ X the observation is fixed!

The Maximum Likelihood Estimator is

$$\hat{\theta} = \underset{\theta \in U}{\operatorname{argmax}}\, f_\theta(X)$$

the point θ which maximizes f_θ
if several, you take one of them.

if I observe $X_1, \dots X_n$ iid with density $g_\theta(x)$

then the density of $X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$ is $f_\theta(X) = g_\theta(X_1) \cdots g_\theta(X_n)$

<u>Notation</u> Likelihood $L(\theta) = f_\theta(X)$ in this case (iid) it is $\prod_{i=1}^{n} g_\theta(X_i)$

log likelihood $\ell(\theta) = \log(f_\theta(X))$

in this case this is $\sum_{i=1}^{n} \log(g_\theta(X_i))$

<u>NB</u> $\hat{\theta}$ is also $\underset{\theta \in \Theta}{\operatorname{argmax}}\, \ell(\theta)$

Kass Eden Brown "Analysis of neural data"

With very few assumptions, MLE are usually

PRO's
- consistant (they cv to θ_0)
- with the smallest asymptotic variance

But

CON's
- you can compute it by hand in very few cases.
- if $L(\theta)$ is computable, its maximisation might be tricky
- there are cases where even computing $L(\theta)$ is a challenge

→ the noise is gaussian ??

$$Y = \text{firing rate} = \frac{\text{Number of spikes}}{\text{Duration of the experiment}}$$

* Number of spikes is usually modeled by Binomial or Poisson
* in both cases when the duration of the experiment is long enough these distribution might be approximated by Gaussian.
* sometimes you need to transform the data to make them look gaussian. The anscombe transform $N \to 2\sqrt{N + 3/8}$ is known to be the best way to make a Poisson look Gaussian.

for the ISI's iid $\mathcal{E}(\lambda)$

$X_1 \dots X_n$ with density $\lambda e^{-\lambda x} \mathbb{1}_{x \geq 0}$.

$$L(\lambda) = f(X) = \prod_{i=1}^{n} \left(\lambda e^{-\lambda X_i} \mathbb{1}_{X_i \geq 0} \right)$$

$$= \lambda^n e^{-\lambda (X_1 + \dots + X_n)} \mathbb{1}_{\min(X_i) \geq 0}$$

$$\ell(\lambda) = n \log(\lambda) - \lambda (X_1 + \dots + X_n) + \cancel{\log\left(\mathbb{1}_{\min(X_i) \geq 0}\right)}$$ is always 0 all $X_i > 0$.

$$\frac{\partial \ell}{\partial \lambda}(\lambda) = \frac{n}{\lambda} - (X_1 + \dots + X_n) \quad \text{null in } \lambda = \left(\frac{X_1 + \dots + X_n}{n} \right)^{-1}$$

	0		$\frac{1}{\overline{X}}$		$+\infty$
$\frac{\partial \ell}{\partial \lambda}$		$+$	0	$-$	
ℓ		$\nearrow$		$\searrow$	

MLE is $\hat{\lambda} = \frac{1}{\overline{X}}$ with $\overline{X} = \frac{X_1 + \dots + X_n}{n}$

In the gaussian model

$$Y_i = \begin{cases} a + b W_i + \sigma \varepsilon_i \\ a + b \cos\theta_i + \sigma \varepsilon_i \end{cases}$$

in general for linear gaussian models (think also to ANOVA etc)

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = m + \sigma \varepsilon \quad \text{with} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad \varepsilon_i \text{ are iid } \mathcal{N}(0,1)$$

and $m \in V \subsetneq \mathbb{R}^n$ (V strict subset of $\mathbb{R}^n$).

m and σ are both unknown

eg $V = \mathrm{Vect}\left(\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \begin{pmatrix} W_1 \\ \vdots \\ W_n \end{pmatrix} \right)$ or $V = \mathrm{Vect}\left(\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \begin{pmatrix} \cos\theta_1 \\ \vdots \\ \cos\theta_n \end{pmatrix} \right)$

$\theta = (\underset{\in V}{\underline{\underline{m}}}, \sigma)$ so $\dim(V) + 1$ parameters

Let us estimate m and σ by MLE.

$$L(\theta) = f_\theta(Y) = \frac{e^{-\frac{(Y_1 - m_1)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \times \dots \times \frac{e^{-\frac{(Y_n - m_n)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$$

$m = \begin{pmatrix} m_1 \\ \vdots \\ m_n \end{pmatrix} \in V \subsetneq \mathbb{R}^n$

in the case $V = \mathrm{Vect}\left(\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \begin{pmatrix} W_1 \\ \vdots \\ W_n \end{pmatrix}\right)$

$m_i = a + bW_i$ $\to$ so just 2 parameters

$$L(m, \sigma) = \frac{\exp\left(-\left(\frac{(Y_1 - m_1)^2 + \dots + (Y_n - m_n)^2}{2\sigma^2}\right)\right)}{\left(\sqrt{2\pi\sigma^2}\right)^n}$$

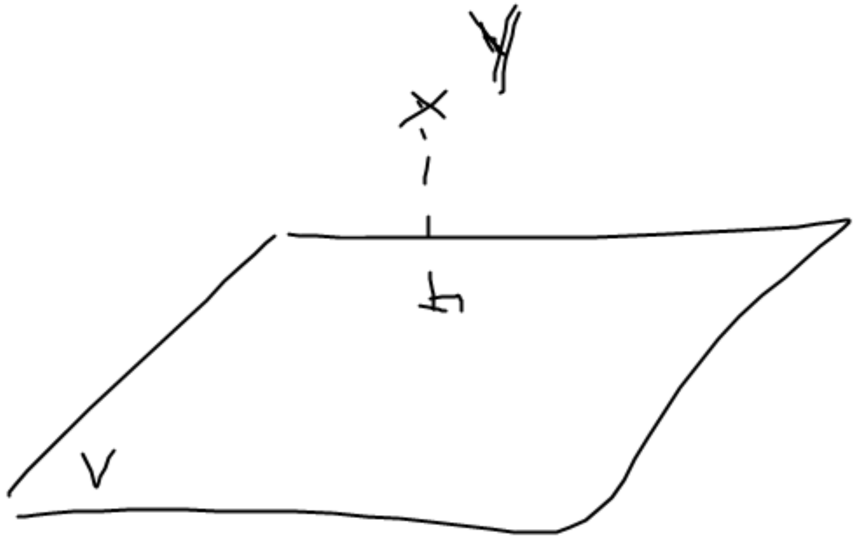
$$= \frac{\exp\left(-\|Y - m\|^2 / 2\sigma^2\right)}{\left(\sqrt{2\pi\sigma^2}\right)^n}$$

$$\ell(m, \sigma^2) = -\frac{\|Y - m\|^2}{2\sigma^2} - \frac{n}{2}\log(2\pi\sigma^2)$$

$$\ell(m,\sigma^2) = -\frac{\|Y-m\|^2}{2\sigma^2} - \frac{n}{2}\log\sigma^2 - \frac{n}{2}\log(2\pi)$$

maximizing the log likelihood will make you minimize

$$\|Y-m\|^2 \quad \text{for } m \in V$$



$\longrightarrow \quad \hat{m} = \Pi_V Y$

exercise if you want:
show that

$$\hat{m} = \Pi_V Y = \begin{pmatrix} \hat{a} + \hat{b} W_1 \\ \vdots \\ \hat{a} + \hat{b} W_n \end{pmatrix}$$

where $\hat{a}, \hat{b}$ are the classical least square estimators..

(when $V = \mathrm{Vect}\left(\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}, \begin{pmatrix}W_1 \\ \vdots \\ W_n\end{pmatrix}\right)$

It remains to maximize in σ^2

$$\ell(\Pi_V(Y), \sigma^2) = \frac{-\|Y - \Pi_V Y\|^2}{2\sigma^2} - \frac{n}{2}\log(\sigma^2)$$

$$\frac{\partial \ell}{\partial \sigma^2} = + \frac{\|Y - \Pi_V Y\|^2}{2(\sigma^2)^2} - \frac{n}{2\sigma^2}$$

$\frac{\partial \ell}{\partial \sigma^2}$ is null in $\widehat{\sigma^2} = \frac{\|Y - \Pi_V Y\|^2}{n}$

<u>Remark</u> $\|Y - \Pi_V Y\|^2 \sim \sigma^2 \chi^2(n - \dim V)$

So $E\left(\|Y - \Pi_V Y\|^2\right) = \sigma^2 (n - \dim V)$

which means that $E(\widehat{\sigma^2}) = \sigma^2 \frac{(n - \dim V)}{n}$

Hence The MLE is biased

$$E(\hat{\sigma}^2_{MLE}) = \sigma^2 \frac{n - \dim V}{n} \neq \sigma^2$$

$$\left(\xrightarrow[n \to +\infty]{} \right)$$

↳ most of the time people prefer $\hat{\sigma}^2_{classic} = \dfrac{\|Y - \Pi_V Y\|^2}{n - \dim V}$

which is unbiased (cf lm( ) in R)

$$E(\hat{\sigma}^2_{classic}) = \frac{\sigma^2 (n - \dim V)}{n - \dim V}$$

$= \sigma^2$ so unbiased.

$$\hat{\sigma}^2_{classic} = \hat{\sigma}^2_{MLE} \times \frac{n}{n - \dim V}$$

Let's go back to the models in section I]

3] a cognitive model of categorization

you are giving to a participant a list of objects

LEARNING						TRANSFER		
								
your answer: A	A	B	A	B	are at random	B	A	B
the teacher: B	A	B	A	B		no feedback		

100%
80%

Models of learning

Models of transfer aim at modeling how a human can categorize given what he has learned (There is no good answer!!)

modeling learning → difficult no independance → ≠ participants

modeling transfer is easier

— you can imagine that the answer for each object is independant of the other ones (no feedback)

- and for the same reason they are "identically distributed" given the object that is presented

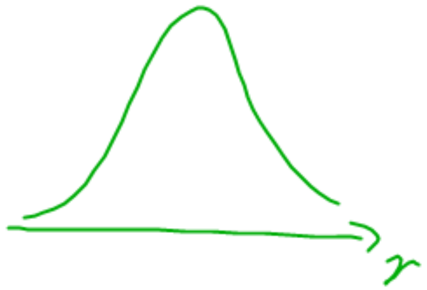
1986, Nosofsky proposed the Generalized Context Model (GCM) for a given object x you represent it by a list of attributes

$$x = (x_1, \dots, x_d)$$

$x_1 \downarrow$ color, $x_d \downarrow$ shape ...

similarity between object x and object y

$$S(x,y) = e^{-c\, d(x,y)}$$



where $d(x,y) = \sum_{i=1}^{d} |x_i - y_i|$

$$\mathbb{P}(y \text{ is said to be in } A) = \frac{\sum_{x \in \mathcal{L} \cap A} S(y,x)}{\sum_{x \in \mathcal{L}} S(y,x)}$$

$$\sum_{x \in \mathcal{L} \cap A} + \sum_{x \in \mathcal{L} \cap B}$$

where $\mathcal{L}$ is the set of learned objects.