Machine Learning for Cognitive Sciences: Principles and Applications

Simona Cocco (Physics Departement of ENS)
simona.cocco@phys.ens.fr CM &TD

Vito Dichio (Physics Departement of ENS)
vito.dichio@phys.ens.fr TD

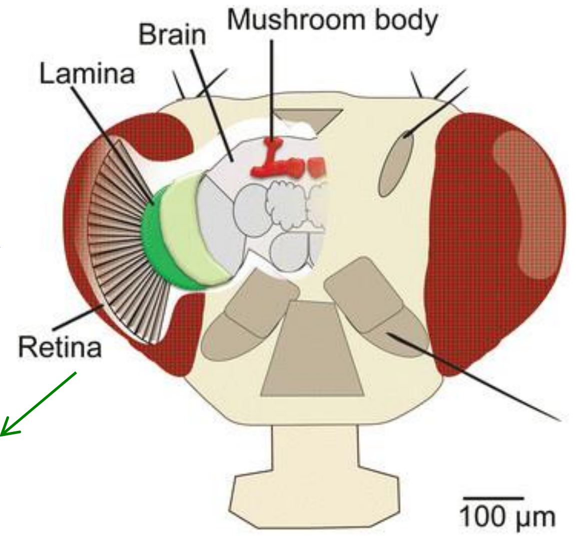
Week 4

Efficient coding

S.B. Laughlin, A Simple Coding Procedure Enhances a Neuron's Information Capacity, 1981

Goal.
explain coding of **light intensity** by the fruit-fly large lamina monopolar cells

Intensity *x*
Distribution *P(x)*



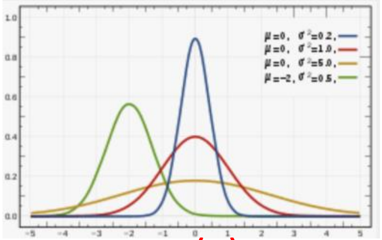
y

Ando et al., 2016

Hp: **Firing rate y** is a Gaussian variable with average value y(x) and variance (σ)

y=y(x)+ weak gaussian noise

$$P(y|x) = \frac{e^{-\frac{(y-y(x))^2}{2\ \sigma^2}}}{\sqrt{2\pi\sigma^2}}$$



y(x)

Optimal code: Maximize MI (x,y),-> Distribution *P(y)* , *Optimal y(x)*

Mutual Information between stimulus Intensity and firing rate.

$$MI(x,y) = \int dx \int dy\, P\,(x,y)\, log_2 \left(\frac{P(x,y)}{P(x)P(y)}\right) = S(P(y)) - S(P(y|x))$$

$$S\big(P(y)\big) = -\int dy\, P(y) \log P(y)$$

$$P(y|x) = \frac{e^{-\frac{(y-y(x))^2}{2\ \sigma^2}}}{\sqrt{2\pi\sigma^2}}$$

- S(P(y|x) $= \int dx\, P(x) \int P(y|x) \log(P(y|x) =$

$=\int dx\, P(x) \int P(y|x) \left(-\frac{(y-y(x))^2}{2\sigma^2} - \frac{1}{2}\, log(2\,\pi\,\sigma^2)\right)$ = - $\frac{1}{2}$ $log_2(2\,\pi\, e\,\sigma^2)$

Entropy of a Gaussian distribution

Mutual Information between stimulus Intensity and firing rate.

$$MI(x,y) = \int dx \int dy\, P\,(x,y)\, log_2\, \left(\frac{P(x,y)}{P(x)P(y)}\right) = -\int dy\, P(y) \log P(y)\ \text{-}\ \frac{1}{2}\, log(2\,\pi\, e\, \sigma^2)$$

Maximise the MI distribution with respect to y: Maximize the Entropy of P(y)

$$\text{Argmax}_{\,P(y)}\, [-\int dy\, P(y) \log P(y) +\ \lambda\, (\int dr\, P(y) - 1)]$$

$$\Downarrow$$

P(y)= const in the interval [0,y_{max}]= $\dfrac{1}{y_{max}}$

Mutual Information between stimulus Intensity and firing rate.

$$MI(x,y) = \int dx \int dy\, P\,(x,y)\, log_2 \left(\frac{P(x,y)}{P(x)P(y)}\right) = -\int dy\, P(y) \log P(y) \;\text{-}\; \frac{1}{2}\, log(2\,\pi\, e\, \sigma^2)$$

Maximise the MI distribution with respect to y: Maximize the Entropy of P(y)

$$\text{Argmax}_{\,P(y)}\, [-\int dy\, P(y) \log P(y) + \lambda\, (\int dr\, P(y) - 1)]$$

$$\Downarrow$$

P(y)= const in the interval [0,y_{max}]= $\dfrac{1}{y_{max}}$

- In the hypothesis of a small gaussian noise: $y \sim y(x)$

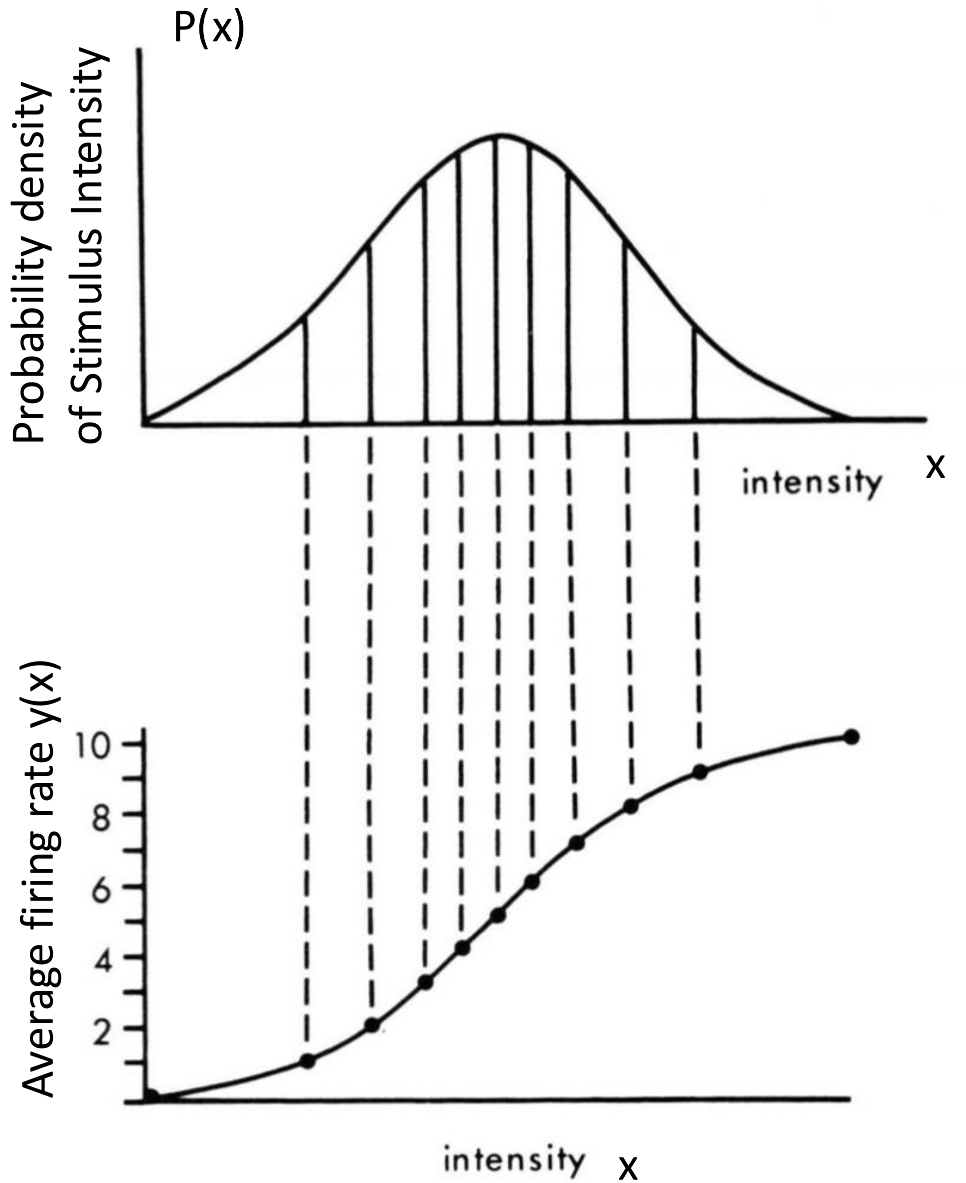
- Looking for *y(x)* such that P(y)=1/y_{max} and P(x) (stimulis intensity) is given.

By changing variable:

$$P(y)\frac{dy}{dx} = P(x) \quad \Rightarrow \quad \frac{y(x)}{y_{max}} = \int_{x_{min}}^{x} P\,(\mathrm{x})\,\mathrm{dx}$$

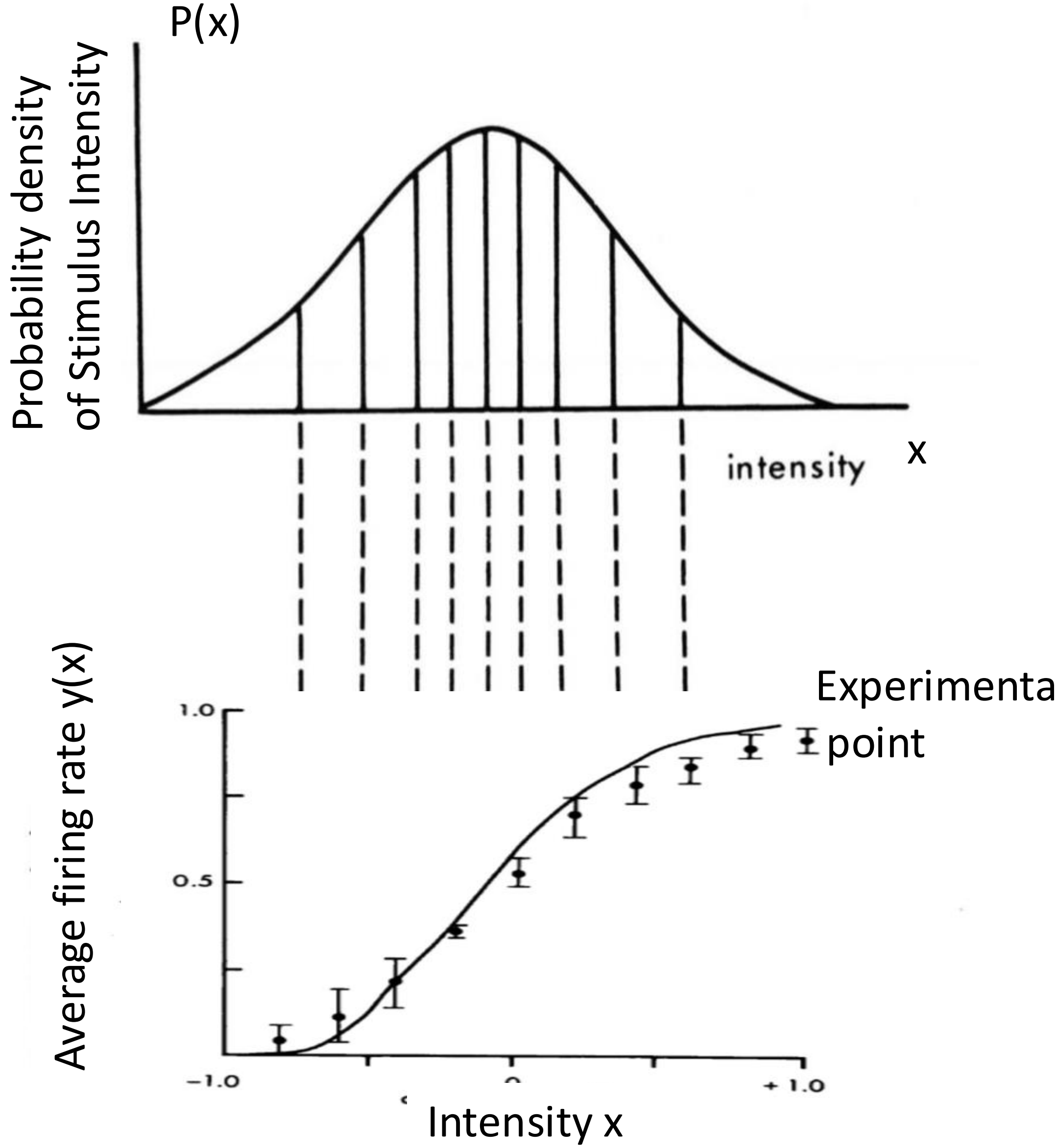
Efficient Coding

$$\frac{y(x)}{y_{max}} = \int_{x_{min}}^{x} P\,(\mathrm{x})\mathrm{dx}$$

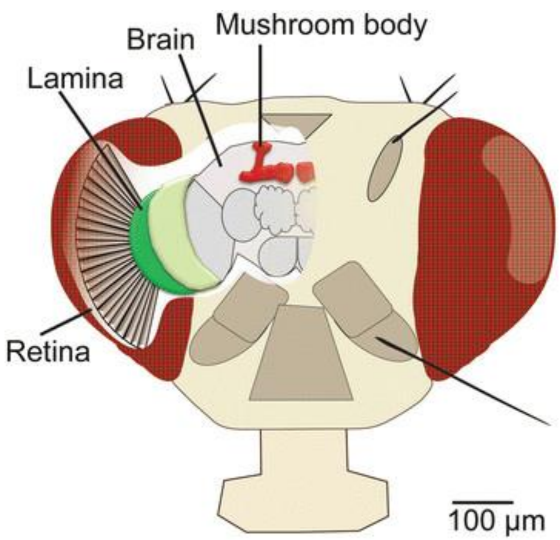


Efficient Coding

S.B. Laughlin, A Simple Coding Procedure Enhances a Neuron's Information Capacity, 1981



$$\frac{y(x)}{y_{max}}=\int_{x_{min}}^{x} P\,(\mathrm{x})\mathrm{dx}$$



Bring home message :

- The Mutual Information between 2 variables quantify the gain in information on one variable knowing the other one

- Efficient Coding: Maximising Mutual Information between Stimulus and firing rate

- The entropy and Asymptotic Inference: Definition of
Cross Entropy and Kullback Leibler Divergence,

Plan of the lecture

Asymptotic Inference and Information Theory.

Definition of Cross Entropy, Kullback Leibler Divergence.

Asymptotic Inference, Entropy & Information Theory

Aim: Have a theoretical understanding of the prediction error :
The difference between the

inferred vecror of parameters and the true one

$$\boldsymbol{\theta} \qquad \hat{\boldsymbol{\theta}},$$

The error vanishes asymptotycally and is controlled by well defined measures in Information Theory

Data vector

$$\boldsymbol{y} \in \mathbb{R}^L$$

Vector of parameters

$$\hat{\boldsymbol{\theta}} \in \mathbb{R}^D,$$

Number of data

$$M$$

$$M \gg D, L$$

Cross Entropy

Consider two distributions $p(\boldsymbol{y})$ and $q(\boldsymbol{y})$.

The **Cross Entropy** is defined as:

$$S_c(p,q) = -\sum_{\boldsymbol{y}} p(\boldsymbol{y}) \log q(\boldsymbol{y}) = -\left\langle \log q \right\rangle_p$$

IF $q(\boldsymbol{y})$ = $p(\boldsymbol{y})$ it coincides with the entropy of $p(\boldsymbol{y})$

Kullback Leibler Divergence

- The KL divergence of $p(\boldsymbol{y})$ with respect to $q(\boldsymbol{y})$ is defined as:

$$D_{KL}(p||q) = \sum_{\boldsymbol{y}} p(\boldsymbol{y}) \log \frac{p(\boldsymbol{y})}{q(\boldsymbol{y})} = S_c(p,q) - S(p)$$

- An important property of the KL divergence is that it is always positive.

- The mutual Information is a KL divergence: tell how much the variable are dependent

$MI(x,y)=D_{KL}(P(x,y)||P(x)P(y))$

Kullback Leibler Divergence

$$S_c(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = S(\hat{\boldsymbol{\theta}}) + D_{KL}(\hat{\boldsymbol{\theta}}||\boldsymbol{\theta})$$

Due to the properties of $D_{KL}(\hat{\boldsymbol{\theta}}||\boldsymbol{\theta})$, the Cross Entropy

enjoys two important properties:

- It is bounded by the Entropy of the true distribution $S_c(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) \geq S(\hat{\boldsymbol{\theta}})$;
- It has a minimum in $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$;

Posterior distribution & Cross Entropy

Consider M data configurations drawn independently:

The likelihood of the data is

$$p(Y|\boldsymbol{\theta}) = \prod_{i=1}^{M} p(\boldsymbol{y}_i|\boldsymbol{\theta}) = \exp\left(M \times \frac{1}{M}\sum_{i=1}^{M}\log p(\boldsymbol{y}_i|\boldsymbol{\theta})\right).$$

The laws of large numbers ensure that:

The true & unknown distribution

$$\frac{1}{M}\sum_{i=1}^{M}\log p(\boldsymbol{y}_i|\boldsymbol{\theta}) \underset{M\to\infty}{\longrightarrow} \int d\boldsymbol{y}\, p(\boldsymbol{y}|\hat{\boldsymbol{\theta}})\, \log p(\boldsymbol{y}|\boldsymbol{\theta}).$$

The inferred distribution

$$p(\boldsymbol{\theta}|Y) \propto p(Y|\boldsymbol{\theta}) \approx e^{-M\, S_c(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})}$$

Convergence of inferred parameters towards their ground-truth value

To obtain the complete expression of the posterior distribution we introduce the Denominator:

$$p(\boldsymbol{\theta}|Y) = \frac{e^{-MS_c(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})}}{\int \mathrm{d}\boldsymbol{\theta}\, e^{-MS_c(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})}} \ .$$

$$p(\boldsymbol{\theta}|Y) \sim e^{-M[S_c(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})-S(\hat{\boldsymbol{\theta}})]} = e^{-MD_{KL}(\hat{\boldsymbol{\theta}}||\boldsymbol{\theta})} \ .$$

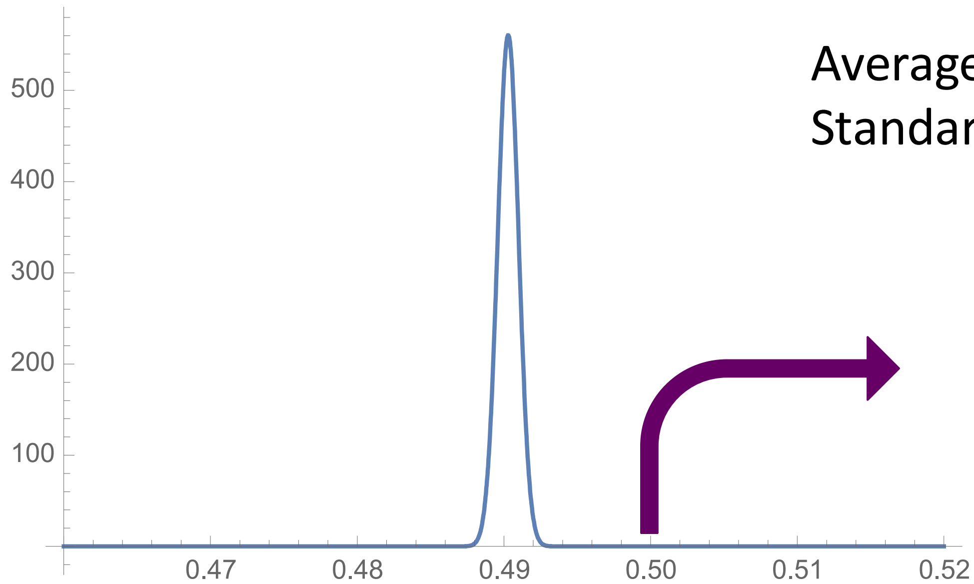
$$D_{KL}(\hat{\boldsymbol{\theta}}||\boldsymbol{\theta}_{hyp})$$

Controls how the posterior probability of the hypotesis θ varies with the number of data: If the hypothesis is not the good one its probability decays exponentially with the number of data.

D_{KL} gives the inverse of number of data you need to realize that your hypothesis is wrong

Laplace and the birth rate of boys & girls

Posterior distribution:



Average θ = 0.490291
Standard deviation θ = 0.007117

Probability that θ exceeds 0.5 = $\int_{0.5}^{1} d\theta \; p(\theta|y) \approx 10^{-42}$ Extremely unlikely!

D_{KL}=42 X log 10 /M $\sim 2\ 10^{-4}$

You need 5 000 data to understand that the probability are not equal

Supplementary Slides

Positivity of Kullback Leibler Divergence

- The KL divergence of $p(\boldsymbol{y})$ with respect to $q(\boldsymbol{y})$ is defined as:

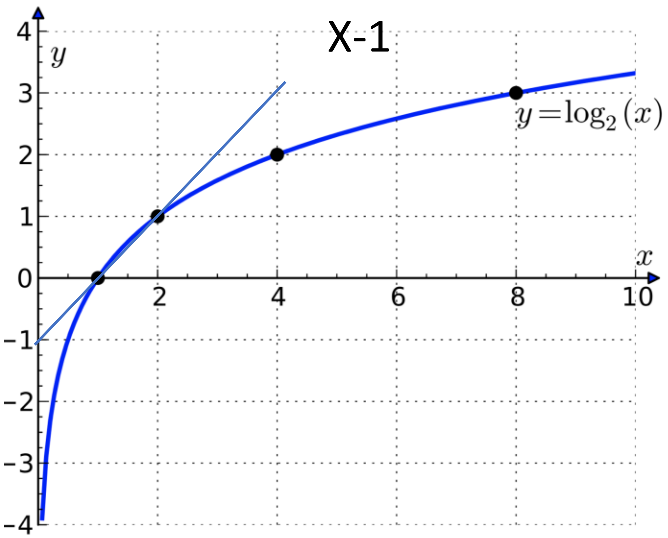
$$D_{KL}(p||q) = \sum_{\boldsymbol{y}} p(\boldsymbol{y}) \log \frac{p(\boldsymbol{y})}{q(\boldsymbol{y})} = S_c(p,q) - S(p)$$

- An importamt property of the KL divergence is that it is always positive:

$$D_{KL}(p||q) \geq 0 \ ,$$

This properties derives from concavity of the logarithm

$$\log x \leq x - 1.$$



- It is zero when the 2 probability distribution coincide:

p=q

- Due to the positivity of the D_{KL} $S_c(p,q) \geq S(p)$