Neuroinformatics Lecture (L10)

Prof. Dr. Gordon Pipa

Institute of Cognitive Science University of Osnabrück

Likelihood

Rehearsal

R

Bayes' Theorem:

posterior ∝ likelihood × prior

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Now use:

y=D (observed data)

x=$\vec{w}$ (set model parameter)

$$p(\vec{w}|D) = \frac{p(D|\vec{w})p(\vec{w})}{p(D)}$$

$$p(D) = \int P(D|\vec{w})p(\vec{w})d\vec{w}$$

Normalisation

Now $p(\vec{w}|D)$ probability of a model with set of parameters given the data

Bayes' Theorem:

Posterior:

probability of a model with set of parameters given the data

$$p(\vec{w}|D)$$

Likelihood:

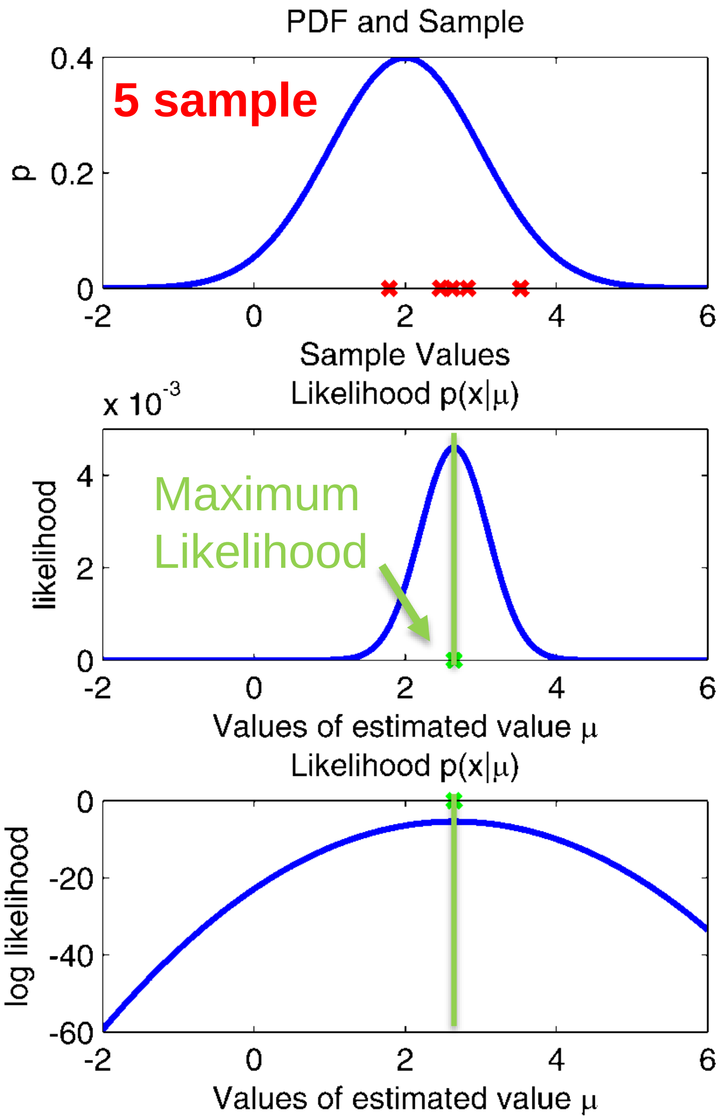
probability of the data given with set of parameters of a model

$$p(D|\vec{w})$$

Likelihood function

Rehearsal

R



Likelihood: $p(D|\vec{w})$

Likelihood function joint probability for all x_n conditioned on the parameters! Here we assume a Gaussian PDF.

$$p(\mathbf{x}|\mu,\sigma^2) = \prod_{i=1}^{N} \mathcal{N}(x_i|\mu,\sigma^2)$$

$$p(\vec{x}|\mu,\sigma) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{(x_i-\mu)^2}{2\sigma^2}\right)}$$

$$p(\vec{x}|\mu,\sigma^2) = f(\mu,\vec{x},\sigma^2) = L(\mu)$$

Matlab code to generate these figures is available on STUDIP: Likelihood_Gauss.m

Maximum Likelihood

Rehearsal

R

Step 1: Write down likelihood: $p(\vec{x}|w_1, w_2, \ldots)$

Step 2: Assume sample are independent and from the same distribution (i.i.d):

$$p(\vec{x}|w_1, w_2, \ldots) = \prod_{i=1}^{N} p(x_i|w_1, w_2, \ldots)$$

Step 3: Assume a certain type of a distribution for example Gaussian

$$p(\vec{x}|\mu, \sigma) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{(x_i-\mu)^2}{2\sigma^2}\right)}$$

Step 4: Maximize this function in respect the wanted parameter

$$\underset{\mu}{\arg\max} \quad p(\vec{x}|\mu, \sigma) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{(x_i-\mu)^2}{2\sigma^2}\right)}$$

$$\underset{\mu}{\arg\max} \ \log p(\vec{x}|\mu, \sigma) = \sum_{i=1}^{N} \left[\log \frac{1}{\sigma\sqrt{2\pi}} - \left(\frac{(x_i - \mu)^2}{2\sigma^2}\right)\right]$$

Step 5: Compute the derivative and set this to zero

$$\frac{\partial}{\partial\mu} \sum_{i=1}^{N} \left[\log \frac{1}{\sigma\sqrt{2\pi}} - \left(\frac{(x_i - \mu)^2}{2\sigma^2}\right)\right] = 0 - \frac{\partial}{\partial\mu} \sum_{i=1}^{N} \left(\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

Maximum Likelihood

Rehearsal

R

- Step 7: set derivative to zero and resort term to get a function $f(x_1, x_2, \dots, x_N | w_2, \dots)$

Maximum = slope flat

$$\frac{\partial}{\partial \mu} \sum_{i=1}^{N} \left[\log \frac{1}{\sigma\sqrt{2\pi}} - \left(\frac{(x_i - \mu)^2}{2\sigma^2} \right) \right] = + \frac{2}{2\sigma^2} \sum_{i=1}^{N} (x_i - \mu) \quad = 0$$

$$0 = + \frac{2}{2\sigma^2} \sum_{i=1}^{N} (x_i - \mu)$$

Lösung 1: σ infinitely large(irrelevant for us)

Lösung 2: $0 = \sum_{i=1}^{N} (x_i - \mu)$

Lösung 2: $0 = \sum_{i=1}^{N} (x_i - \mu^{ML}) \quad \Leftrightarrow 0 = -N\mu^{ML} + \sum_{i=1}^{N} x_i$

$$\mu^{ML} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Summary: Bernoulli Distribution

Distribution: $p(x|\mu) = \mu^{x}(1-\mu)^{1-x}$

Expected value: $E[x] = \mu$

Variance: $var[x] = \mu(1-\mu)$

```
Y = binopdf(x,n,p)
```

For n=1, computes the Bernoulli pdf at each of the values in X using the corresponding probability P. The values in P must lie on the interval [0, 1].

ML for Bernoulli

Given: $\mathcal{D} = \{x_1, \ldots, x_N\}$, m heads (1), $N - m$ tails (0)

likelihood:

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} \ln p(x_n|\mu) = \sum_{n=1}^{N} \{x_n \ln \mu + (1-x_n)\ln(1-\mu)\}$$

Partial derivative after μ (for ML set to zero):

$$\frac{\partial}{\partial \mu} \ln p(\vec{x}\,|\mu) = \sum_{n=1}^{N} \left(x_n \frac{\partial}{\partial \mu} \ln \mu + (1-x_n) \frac{\partial}{\partial \mu} \ln(1-\mu) \right)$$

ML for Bernoulli

Given: $\mathcal{D} = \{x_1, \ldots, x_N\}$, m heads (1), $N - m$ tails (0)

Partial derivative after µ (for ML set to zero):

$$\frac{\partial}{\partial \mu} \ln p(\vec{x}\,|\mu) = \sum_{n=1}^{N} \left(x_n \frac{\partial}{\partial \mu} \ln \mu + (1 - x_n) \frac{\partial}{\partial \mu} \ln(1 - \mu) \right)$$

$$0 = \sum_{n=1}^{N} \left(x_n \frac{1}{\mu} + (1 - x_n) \frac{-1}{1 - \mu} \right)$$

$$0 = \sum_{n=1}^{N} (x_n (1 - \mu) - (1 - x_n) \mu)$$

$$0 = \sum_{n=1}^{N} (x_n - \mathrm{x}_n \mu - \mu + x_n \mu)$$

ML for Bernoulli

Given: $\mathcal{D} = \{x_1, \ldots, x_N\}$, m heads (1), $N-m$ tails (0)

Partial derivative after µ (for ML set to zero):

$$\frac{\partial}{\partial\mu} \ln p(\vec{x}\,|\mu) = \sum_{n=1}^{N} \left(x_n \frac{\partial}{\partial\mu} \ln\mu + (1-x_n)\frac{\partial}{\partial\mu}\ln(1-\mu) \right)$$

$$0 = \sum_{n=1}^{N} (x_n - \mathrm{x}_n\mu - \mu + x_n\mu)$$

$$0 = \sum_{n=1}^{N} (x_n - \mu) = -N\mu + \sum_{n=1}^{N} x_n$$

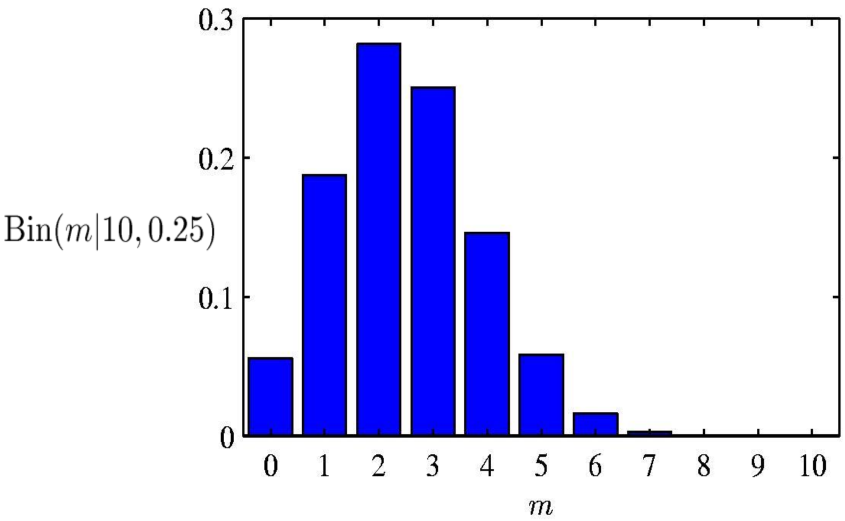
$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

ML estimator of $p(x_n=1)$, with $p(x_n=0)=1-p(x_n=1)$

Next Topic: ML for the expected value of a Binominal distribution

Binomial Distribution

$$\mathrm{Bin}(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m}$$



$$\mathbb{E}[m] \equiv \sum_{m=0}^{N} m\mathrm{Bin}(m|N,\mu) \quad = \quad N\mu$$

$$\mathrm{var}[m] \equiv \sum_{m=0}^{N} (m - \mathbb{E}[m])^2\, \mathrm{Bin}(m|N,\mu) \quad = \quad N\mu(1-\mu)$$

Y = binopdf(x,n,p)

computes the binomial pdf at each of the values in X using the corresponding parameters in N and P. The parameters in N must be positive integers, and the values in P must lie on the interval [0, 1].

ML for Binominal

Given: $\mathcal{D} = \{x_1, \dots, x_N\}$, m heads (1), $N - m$ tails (0)

likelihood: $p(m|\mu, N) = \binom{N}{m} \cdot \mu^m (1-\mu)^{N-m}$

Log likelihood: $\log p(m|\mu, N) = \log\left(\binom{N}{m} \cdot \mu^m (1-\mu)^{N-m} \quad \right)$

$$\log p(m|\mu, N) = \log\binom{N}{m} + \mathrm{m} \log \mu + (N-m)\log(1-\mu)$$

Partial derivative after µ (for ML set to zero):

$$\frac{\partial}{\partial\mu} \ln p(m\,|\mu, N) = \frac{m}{\mu} - \frac{N-m}{1-\mu} = 0$$

$$m(1-\mu) - \mu(N-m) = m - m\mu - \mu N + m\mu = 0$$

$$\mu_{ML} = \frac{m}{N}$$

ML estimator of $p(x_n=1)$

ML for Binominal

Given: $\mathcal{D} = \{x_1, \ldots, x_N\}$, m heads (1), $N - m$ tails (0)

likelihood: $p(m|\mu, N) = \binom{N}{m} \cdot \mu^m (1 - \mu)^{N-m}$

Derive the log-likelihood and the ML estimator of μ

Rules you may want to use

$$ln \prod x = \sum \ln(x) \qquad \ln(x \cdot y \cdot z) = \ln(x) + \ln(y) + \ln(z) \qquad \ln(a^x) = x\ln(a)$$

Timer (8min): Start Stop

ML for Binomial

Given: $\mathcal{D} = \{x_1, \ldots, x_N\}$, m heads (1), $N - m$ tails (0)

likelihood: $p(m|\mu, N) = \binom{N}{m} \cdot \mu^m (1-\mu)^{N-m}$

Log likelihood: $\log p(m|\mu, N) = \log\left(\binom{N}{m} \cdot \mu^m (1-\mu)^{N-m} \right)$

$$\log p(m|\mu, N) = \log\binom{N}{m} + \mathrm{m}\log\mu + (N-m)\log(1-\mu)$$

Partial derivative after μ (for ML set to zero):

$$\frac{\partial}{\partial\mu}\ln p(m\,|\mu, N) = \frac{m}{\mu} - \frac{N-m}{1-\mu} = 0$$

$$m(1-\mu) - \mu(N-m) = m - m\mu - \mu N + m\mu = 0$$

$$\mu_{ML} = \frac{m}{N}$$

ML estimator of $p(x_n=1)$

Next Topic: General Form of the Likelihood of a Gaussian

Likelihood (i.i.d. normal distributed)

Log-Likelihood

$$\ln p(x|\mu,\sigma^2) = ln \prod_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$= \sum_i ln \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$= \underbrace{\textstyle\sum_i ln \frac{1}{\sigma\sqrt{2\pi}}}_{} + \textstyle\sum_i \ln e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$-N \cdot ln\left(\sqrt{\sigma^2 2\pi}\right)$$

$$= -N \cdot ln\left((\sigma^2)^{\frac{1}{2}}\right) \quad - N \cdot ln\left(\sqrt{2\pi}\right)$$

$$= -\frac{N}{2} \cdot ln(\sigma^2) \quad - \frac{N}{2} \cdot ln(2\pi)$$

$$p(\vec{x}|\mu,\sigma) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{(x_i-\mu)^2}{2\sigma^2}\right)}$$

Rule

$$ln \prod x = \sum \ln(x)$$

$$\ln(x \cdot y \cdot z) = \ln(x) + \ln(y) + \ln(z)$$

$$\ln(a^x) = x\ln(a)$$

Likelihood (i.i.d. normal distributed)

Log-Likelihood

$$\ln p(x|\mu,\sigma^2) = ln \prod_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$= \sum_i ln \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$= \underbrace{\textstyle\sum_i ln \frac{1}{\sigma\sqrt{2\pi}}}_{-\frac{N}{2}\cdot ln(\sigma^2) \quad -\frac{N}{2}\cdot ln(2\pi)} + \textstyle\sum_i \ln e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$p(\vec{x}|\mu,\sigma) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{(x_i-\mu)^2}{2\sigma^2}\right)}$$

Rule

$$ln \prod x = \sum \ln(x)$$

$$\ln(x \cdot y \cdot z) = \ln(x) + \ln(y) + \ln(z)$$

$$\ln(a^x) = x\ln(a)$$

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

Next Topic: ML for the variance of a normal distributed random variable

Maximum (Log) Likelihood

Maximum Likelihood for Variance (i.i.d. normal distributed)

$$p(\vec{x}|\mu,\sigma) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{(x_i-\mu)^2}{2\sigma^2}\right)}$$

Using Log trick to get general form of the Log likelihood of a Gaussian

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

$$\frac{\partial}{\partial\sigma}\ln p(\overrightarrow{x\,|}\,\mu,\sigma^2) = \frac{-1(-2)}{2\sigma^3}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{\sigma}$$

Maximum (Log) Likelihood

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

Score function (for ML we set the score function to zero):

$$\frac{\partial}{\partial\sigma}\ln p(\overrightarrow{x} \mid \mu,\sigma^2) = \frac{-1(-2)}{2\sigma^3}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{\sigma}$$

$$0 = \frac{1}{{\sigma_{ML}}^3}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{\sigma_{ML}}$$

$$0 = \frac{1}{\sigma_{ML}}\left(\frac{1}{{\sigma_{ML}}^2}\sum_{n=1}^{N}(x_n-\mu)^2 - N\right)$$

Maximum (Log) Likelihood

Maximum Likelihood for expected value (i.i.d. normal distributed)

$$0 = \frac{1}{\sigma_{ML}}\left(\frac{1}{{\sigma_{ML}}^2}\sum_{n=1}^{N}(x_n-\mu)^2 - N\right)$$

Solution 1: Variance is infinite.

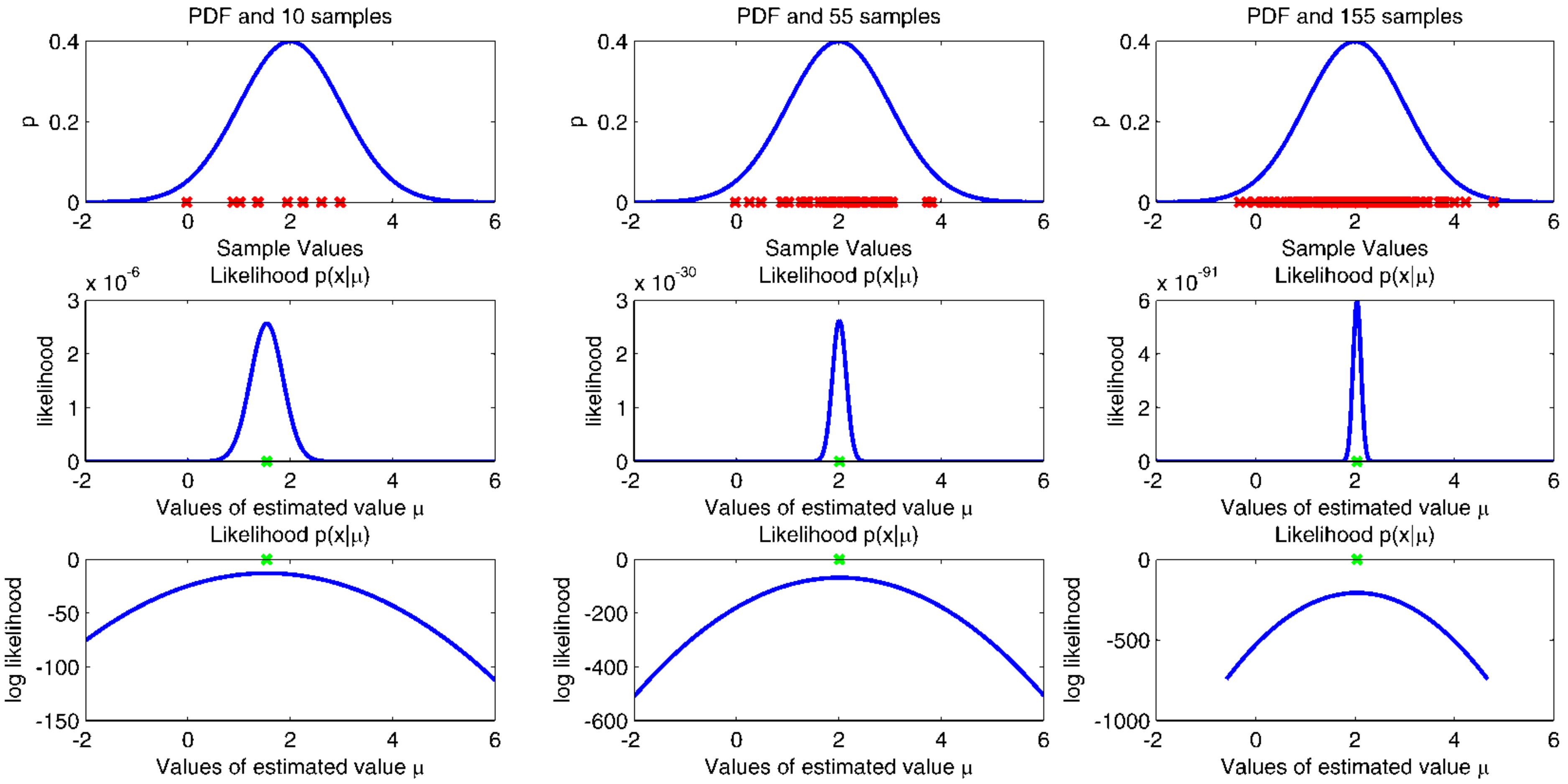
Solution 2: $$0 = \frac{1}{{\sigma_{ML}}^2}\sum_{n=1}^{N}(x_n-\mu)^2 - N$$

$$N = \frac{1}{{\sigma_{ML}}^2}\sum_{n=1}^{N}(x_n-\mu)^2$$

$${\sigma_{ML}}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^2$$

ML estimator of variance

Next Topic: Fischer Information



With the increasing number of samples, the likelihood function becomes tighter and the curvature of the log likelihood increases.

Matlab code to generate these figures is available on STUDIP: Likelihood_Gauss.m

Fischer Information

Fisher Information measures the importance of a parameter for the model

Score function: $$S(w) = \frac{\partial}{\partial w} log L(\vec{x}, w)$$

ML: $$S(w_{ML}) = 0$$

Fischer Information: $$I(w) = -\frac{\partial^2}{\partial w^2} log L(\vec{x}, w)$$

Measures the curvature. For w=w_{ML} , it measures the curvature at the maximum likelihood estimators of w.

A tight and high peak indicates high sensitivity toward that parameter. Therefore the Fischer information is large.

Fischer Information

Fischer Information:
(Expected Fischer Info.)

$$I(w) = -\frac{\partial^2}{\partial w^2} log L(\vec{x}, w)$$

- The Fisher information is a way of measuring the amount of information that an observable random variable X carries about an unknown parameter θ upon which the probability of X depends.
- The Fischer information is high if the likelihood changes very strongly with small changes of the parameter θ.
- In turn that means that a large Fischer information indicates a rather precise estimate of the parameter θ

Maximum (Log) Likelihood

So far: Maximum Likelihood for expected value (normal distributed)

Than maximize the likelihood after w: $p(D|\vec{w}) = \prod_i N(x_i|\mu, \sigma)$

Score function: $S(w) = \frac{\partial}{\partial w} log L(\vec{x}, w)$

For the expected value of a Gaussian

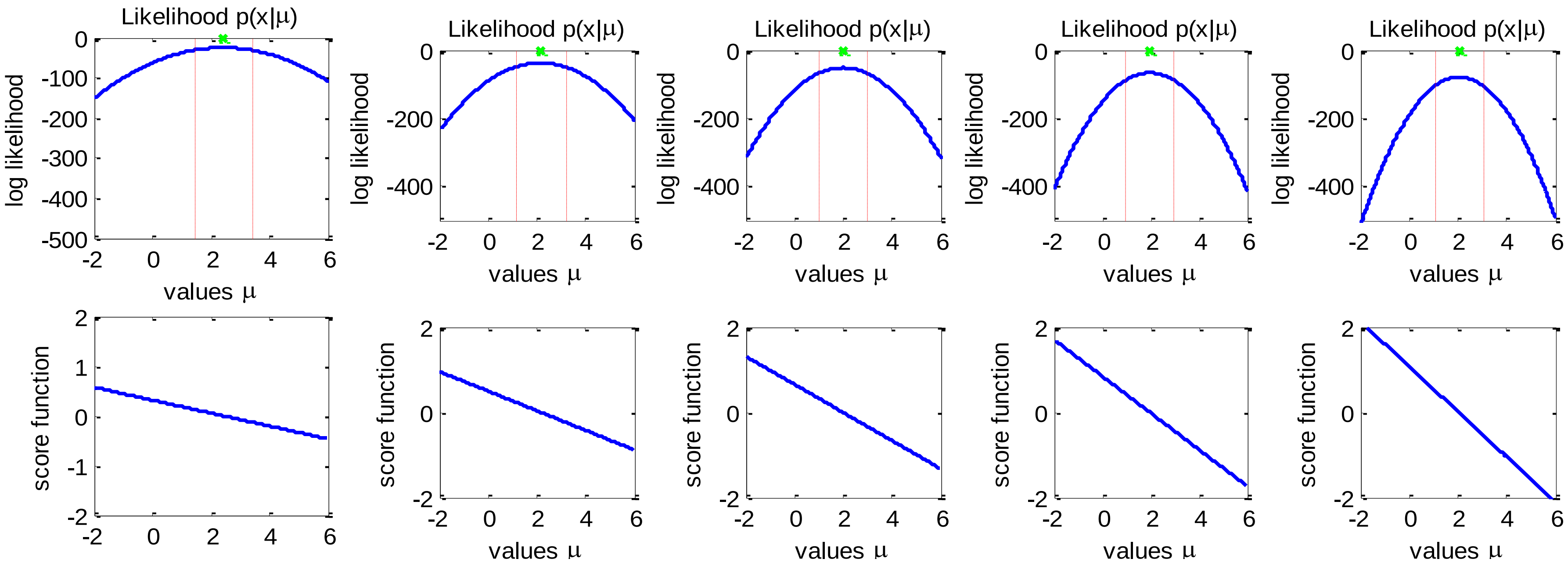
$$S(\mu) = \frac{\partial}{\partial \mu} \log p(\vec{x}|\mu, \sigma) = +\frac{2}{2\sigma^2} \sum_{i=1}^{N} (x_i - \mu)$$

$$S(\mu) = +\frac{1}{\sigma^2} \sum_{i=1}^{N} x_i \quad -\frac{N\mu}{\sigma^2}$$

Given a set of data points this is a straight line with slope N (number of data points)

Fischer Information

Increasing Fischer Information



Maximum (Log) Likelihood

Score function: $S(w) = \frac{\partial}{\partial w} logL(\vec{x}, w)$ $S(\mu) = +\frac{1}{\sigma^2}\sum_{i=1}^{N} x_i \quad -\frac{N\mu}{\sigma^2}$

Fischer Information: $I(w) = -\frac{\partial^2}{\partial w^2} logL(\vec{x}, w)$

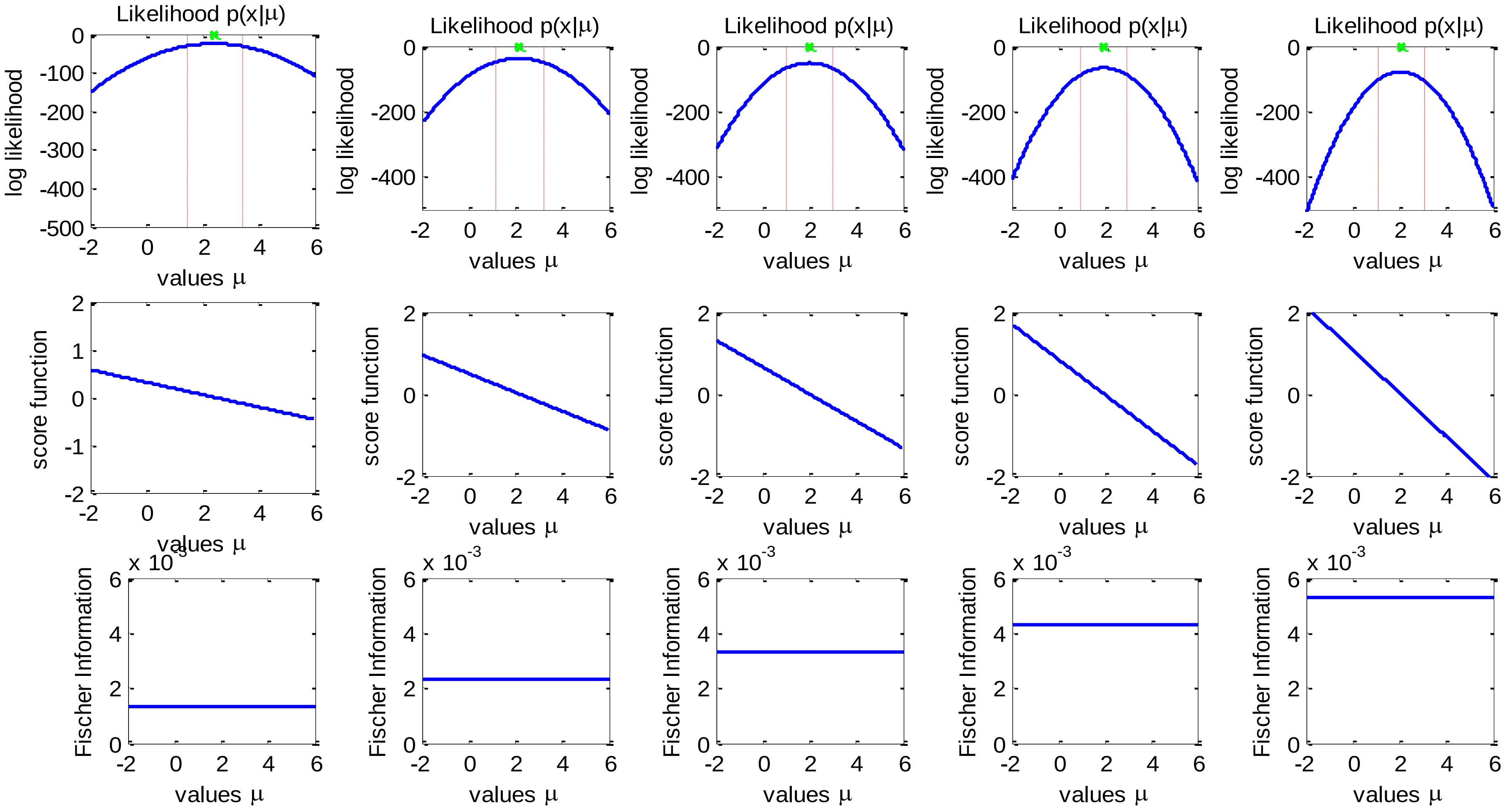
For the expected value of a Gaussian

$$I(\mu) = \frac{\partial^2}{\partial \mu^2}\log p(\vec{x}|\mu,\sigma) = \frac{\partial}{\partial \mu} S(\mu) = \frac{\partial}{\partial \mu}\left(\frac{1}{\sigma^2}\sum_{i=1}^{N} x_i \quad -\frac{N\mu}{\sigma^2}\right) = 0 + \frac{N}{\sigma^2}$$

A constant = N (number of data points) divided by variance

Fischer Information

Increasing Fischer Information



Interpretation : Fischer Information

For the expected value of a Gaussian

$$I(\mu) = +\frac{N}{\sigma^2}$$

Interpretation:

- With increasing number of samples the Fischer information increases ⬄ We know more about the paramter µ ⬄ Higher precision of the estimation
- The larger the variance the lower the Fischer Information ⬄ We know less about the paramter µ ⬄ lower precision of the estimation

Neuroinformatics Questionnaire

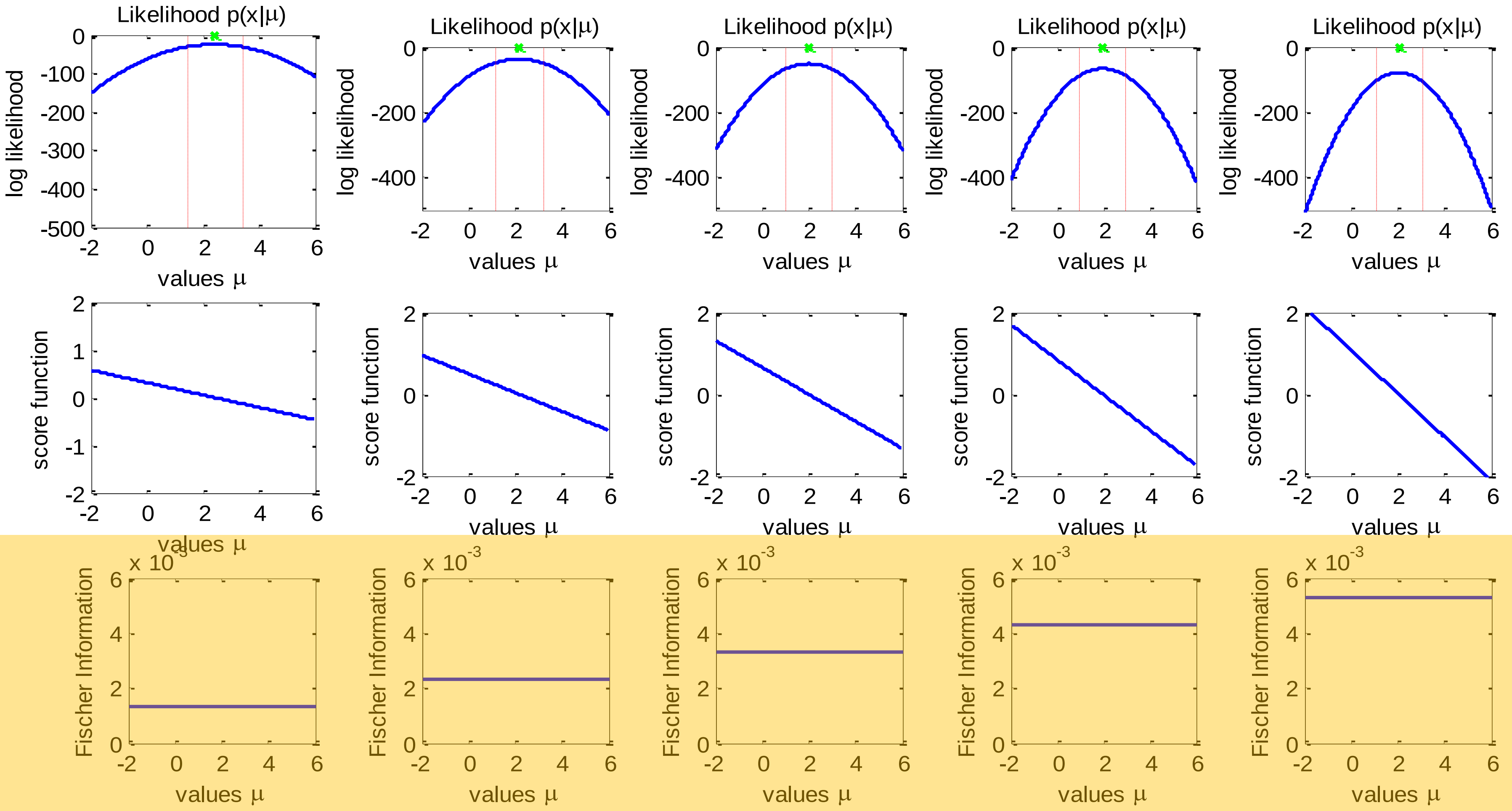
4. Write down the definition of the Fischer Information. Explain what it measures.

Timer (5min): Start Stop

- The Fisher information is a way of measuring the amount of information that an observable random variable X carries about an unknown parameter q upon which the probability of X depends.

- The Fischer information is high if the likelihood changes very strongly with small changes of the parameter q.

- In turn that means that a large Fischer information indicates a rather precise estimate of the parameter q


Likelihood p(x|μ)
log likelihood
0
-100
-200
-300
-400
-500
-2
0
2
4
6
values μ
score function
2
1
0
-1
-2
Fischer Information
x 10^-3
6
4
2
0

4W

The 4W question session

What is this it all about ?

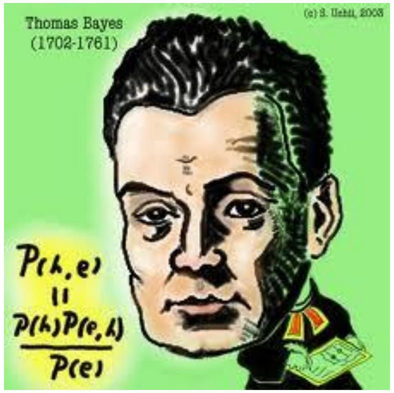
What can I use it for ?

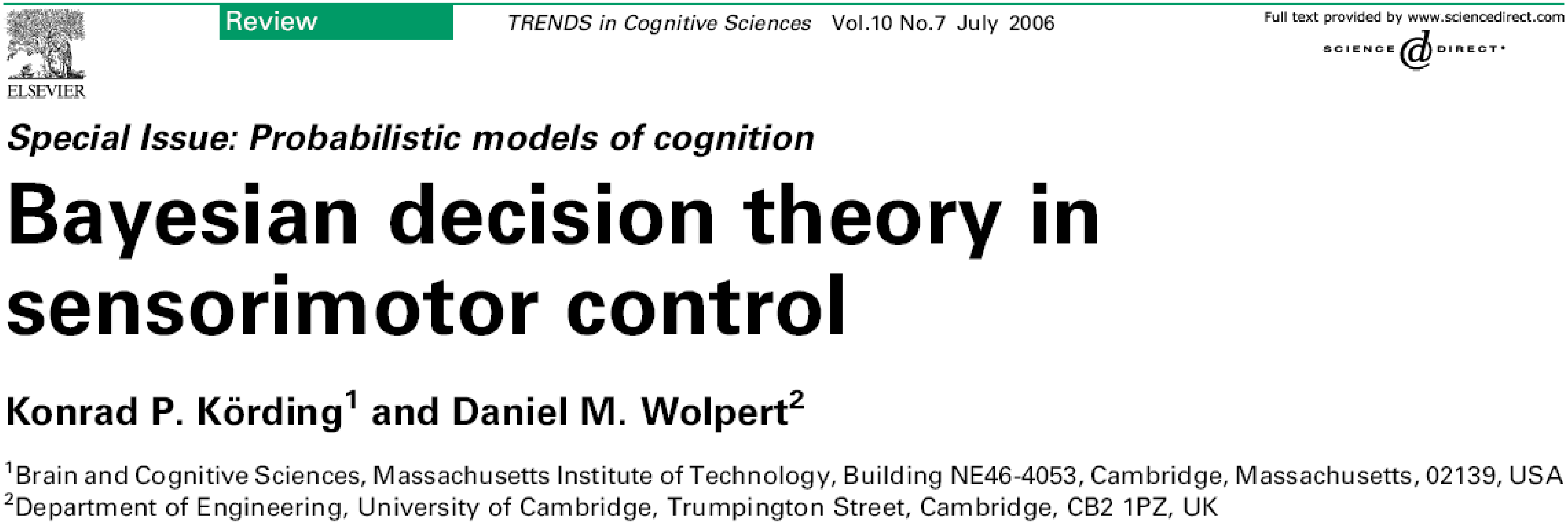
Why should I learn it ?

What are potential Bachelor and Master thesis topics ?

The Brain is Bayesian

4W





Review

TRENDS in Cognitive Sciences Vol.10 No.7 July 2006

Full text provided by www.sciencedirect.com

SCIENCE DIRECT

ELSEVIER

Special Issue: Probabilistic models of cognition

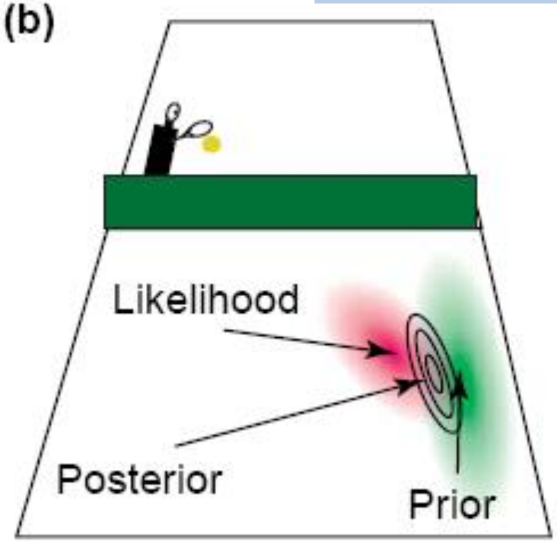
Bayesian decision theory in sensorimotor control

Konrad P. Körding[1] **and Daniel M. Wolpert**[2]

[1] Brain and Cognitive Sciences, Massachusetts Institute of Technology, Building NE46-4053, Cambridge, Massachusetts, 02139, USA
[2] Department of Engineering, University of Cambridge, Trumpington Street, Cambridge, CB2 1PZ, UK

Paper is uploaded in Studip

4W



Decision theory quantifies how people should choose in the context of a given utility function and some partial knowledge of the world. The expected utility is defined as:

$$E[Utility] \equiv \sum_{\substack{possible\\outcomes}} p(outcome|action)U(outcome)$$

where $p(outcome|action)$ is the probability of an outcome given an action and $U(outcome)$ is the utility associated with this outcome.

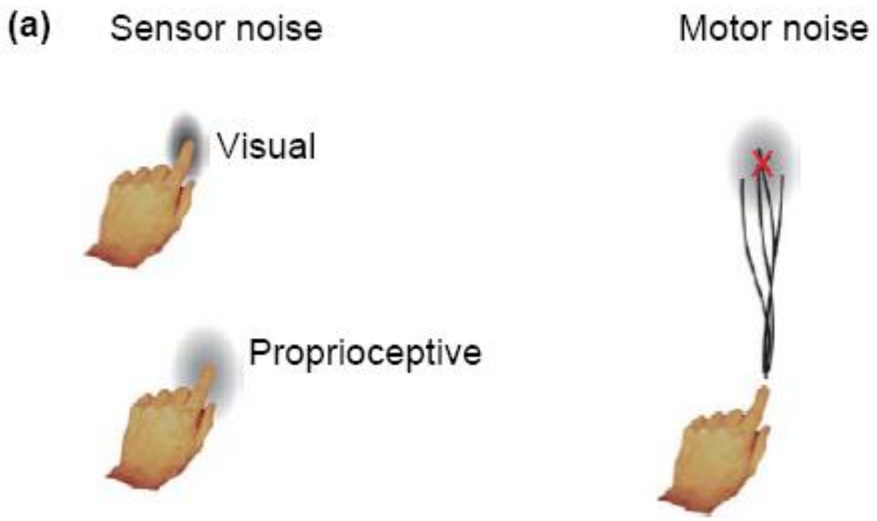
Outcome defined by:

- Prior of ball position
- Likelihood of ball position
- Prior of success in making a point dependent on the position

Is the brain really Bayesian ?

Experiment :

- A subject needs to point to a target.
- There are three types of noise leading uncertainty



4W

Box 2. Bayesian statistics

When we have a Gaussian prior distribution $p(x)$ and we have a noisy observation o of the position that leads to a Gaussian likelihood (red curve, Figure I) $p(o|x)$ it is possible to use Bayes rule to calculate the posterior distribution (yellow curve, Figure I; how probable is each value given both the observation and the prior knowledge):

$$p(x|o) = p(o|x)\frac{p(x)}{p(o)}$$

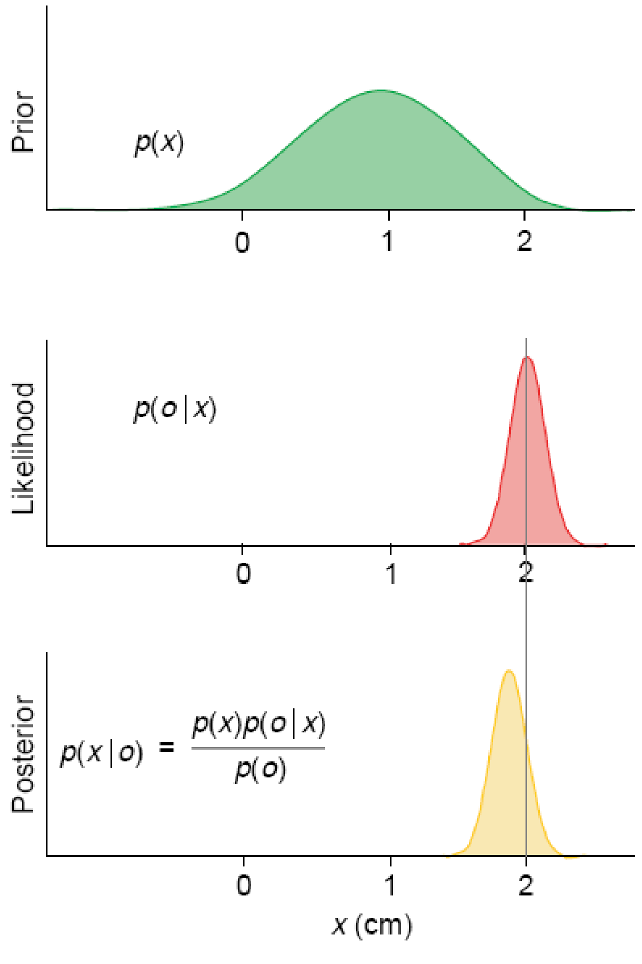
This equation assigns a probability to every possible location. If we assume that the prior distribution $p(x)$ is a symmetric one dimensional Gaussian with variance σ_{p}^2 and mean $\hat{\mu}$ and that the likelihood $p(o|x)$ is also a symmetric one dimensional Gaussian with variance σ_o^2 and mean o, it is possible to compute the posterior that is then also Gaussian in an analytical way. The optimal estimate $\hat{x}$, that is the maximum of the posterior is:

$$\hat{x} = \alpha o + (1 - \alpha)\hat{\mu}$$

where

$$\alpha = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_o^2}$$

Moreover we can calculate the width of the posterior as $\sigma^2 = \alpha\sigma_o$. The parameter α is always less than 1. This Bayesian approach leads to a better estimate of possible outcomes than any estimate that is only based on the sensory input.



TRENDS in Cognitive Sciences

Figure I. Bayesian integration. The green curve represents the prior and red curve represents the likelihood. The yellow curve represents the posterior, the result from combining prior and likelihood.

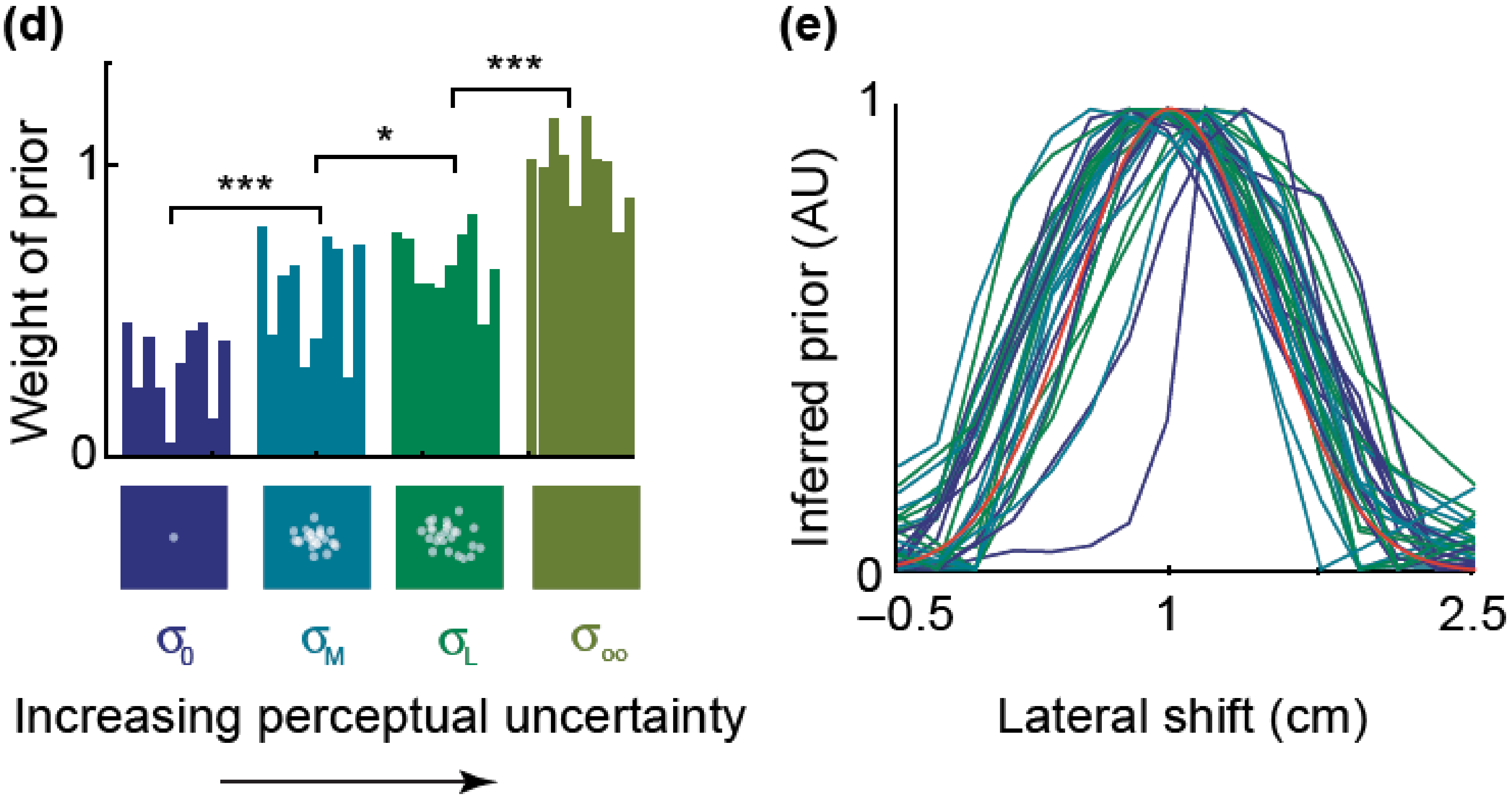
Is the brain really Bayesian ?

Experiment :

- A subject needs to point to a target.
- There are three types of noise leading uncertainty

Test:

- Can the subject learn the uncertainty of the indiviudal sources to estimate the width of the likelihood?
- Can the subject learn a prior?
- Can it combine the prior and likelihood such that it makes a decision that is maximizing the posterior?



Yes we can, actually for this example it is Baysian optimal!