Parameter Estimation
MLE vs. Bayesian (MAP)

In the Bayesian approach, we consider parameters as random variables with a distribution allowing us to model our uncertainty in estimating them.

Ethem Alpaydin, "Intro to ML"

Nagiza F. Samatova, samatova@csc.ncsu.edu
Professor, Department of Computer Science
North Carolina State University

Senior Scientist, Computer Science & Mathematics Division
Oak Ridge National Laboratory

Outline

- Frequentist vs. Bayesian View on Parameter Estimation
  - Benefits of Bayesian Parameter Estimation
- Likelihood & Log-Likelihood Functions
  - Primer: Joint Probability Distribution for i.i.d. sample
  - Example: Likelihood for a Gaussian distribution
  - Likelihood vs. Log-likelihood
- MLE: Maximum Likelihood Estimation
  - Problem Statement
  - Prediction with MLE Parameter Estimators
  - Algebraic, Analytic, Numeric Solutions for MLE
  - MLE vs. OLS
- Bayesian Parameter Estimation
  - MAP
  - Prediction with Parameter Bayesian Estimators
  - Bayesian Regression

Diachronic Interpretation of Bayes Theorem

H: Hypothesis
E: Evidence

prior beliefs before seeing the evidence

likelihood of observing the evidence if H is correct

$$\boxed{P(H \mid E) = \frac{P(H)\, P(E \mid H)}{P(E)}}$$

posterior probability

likelihood of the evidence under any circumstances; normalizing *constant*

Diachronic means **<u>through time</u>**:

- P(H | E): What is the probability of my hypothesis given that I have seen some new evidence, or
- **if you see some new evidence, then you can update your belief in your hypothesis**

Informally, Frequentist vs. Bayesian

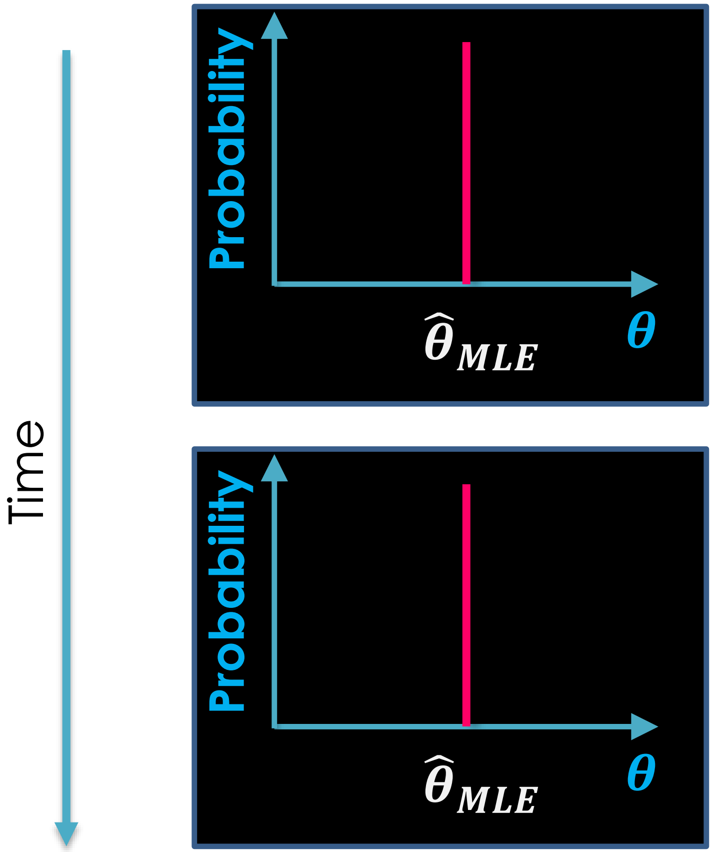
Frequentist: Sampling is infinite, decision rules can be sharp. Data is a repeatable random sample - there is a frequency. Underlying **parameters are fixed**, i.e. they remain **constant** during this repeatable sampling process.

Bayesian: Unknown quantities are treated probabilistically and the state of the world can always be updated. Data are observed from the realized sample. **Parameters are** unknown and described **probabilistically**. It is the data that is fixed.

Src: http://stats.stackexchange.com/questions/22/bayesian-and-frequentist-reasoning-in-plain-english

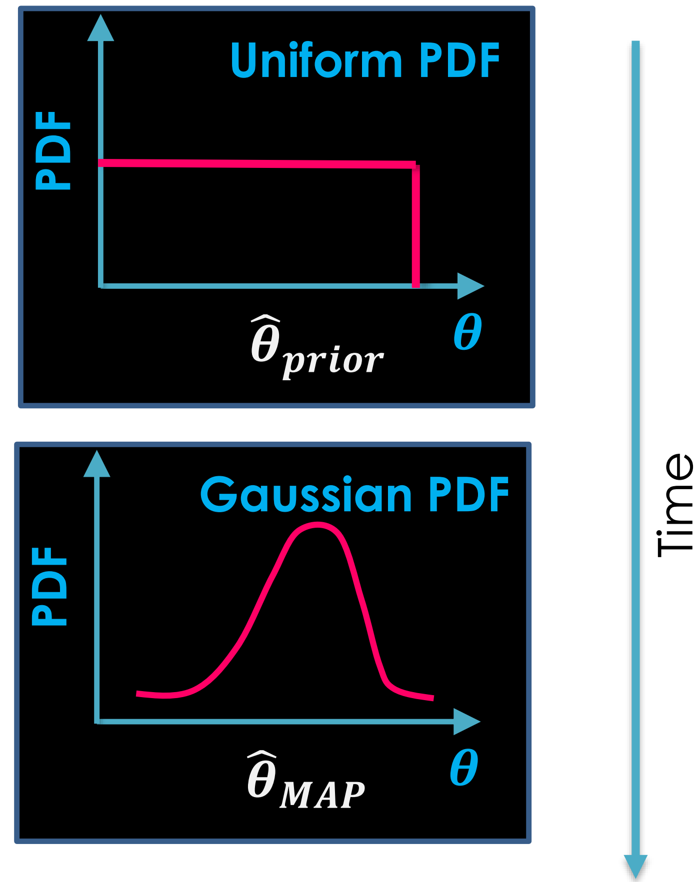
Visually, Frequentist vs. Bayesian

Frequentist's View Point



- Parameter, θ is an unknown **constant**

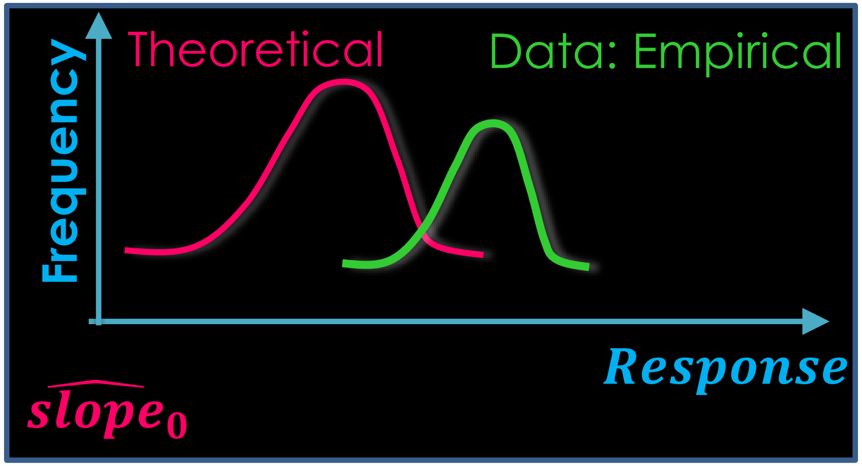
Bayesian View Point

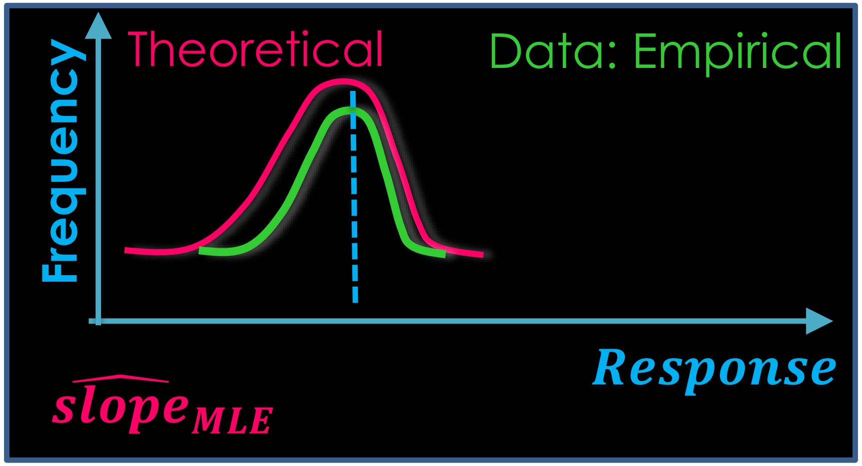


- Parameter, θ is a **random variable** with a probability distribution

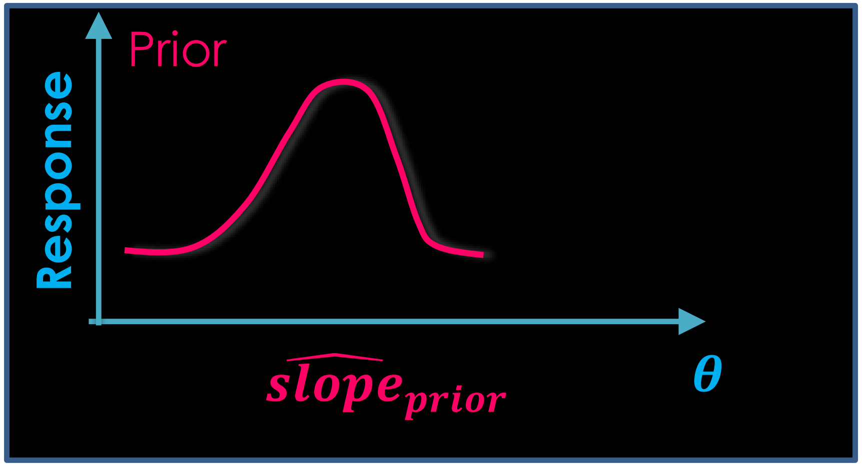
Example: MLE

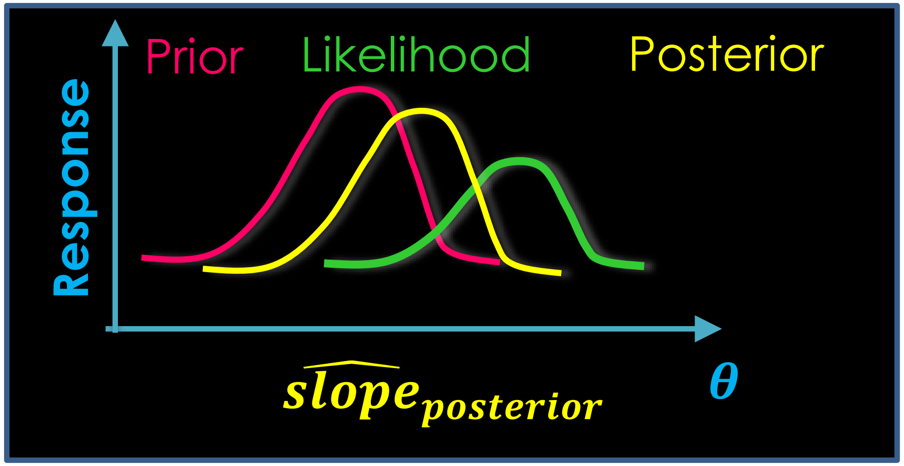
- **MLE Example**: In Linear Regression
  - We estimate the most likely value for the slope and intercept parameters
    - **how well** $slope_{MLE}$ and $intercept_{MLE}$ **fit** the given **data**
  - Make a single prediction for the most likely response value as specified by ($slope_{MLE}$ and $intercept_{MLE}$)





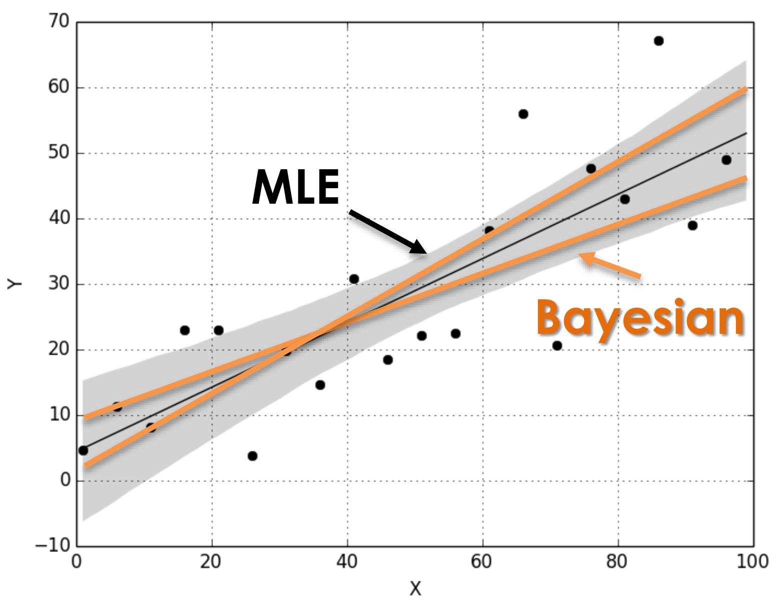
Example: Bayesian





Bayesian Linear Regression Example:

- We can define a prior distribution on the slope and intercept parameters
- Calculate a posterior on them, i.e., distribution over lines
- Average over the prediction of all possible lines weighted by how likely they are as specified by (**weight ~ prior * likelihood**):
  - their prior weights (priors) and
  - how well they fit the given data (likelihoods)



Bayesian Parameter Estimation: Advantages

- **Parameter Search Optimization**: The prior helps ignore the values that parameter θ is unlikely to take
  - To concentrate on the region where it is likely to lie
  - Even a weak prior with long tails can be very helpful

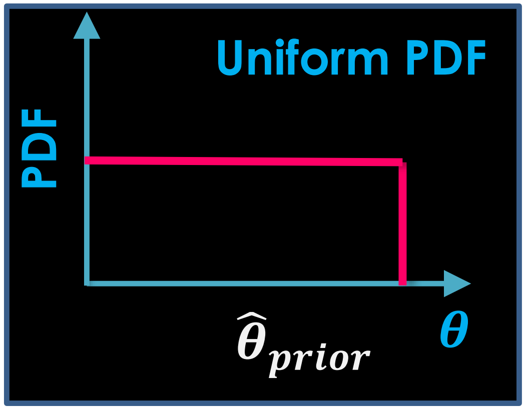
- **Prediction:** Instead of using a single θ estimate in prediction, a set of possible θ values is generated as defined by the posterior
  - To use all of them in prediction,
  - Weighted by how likely each of the value is (i.e., sum or integrate)

- **With θ_{MLE} estimate, we loose both advantages!**
- **With uninformative (uniform) prior, we benefit Prediction but not Parameter Search**

Model-based View on Bayesian Inference

prior beliefs about model parameters: pre-experimental knowledge of parameter values

likelihood of obtaining this data given our choice of θ

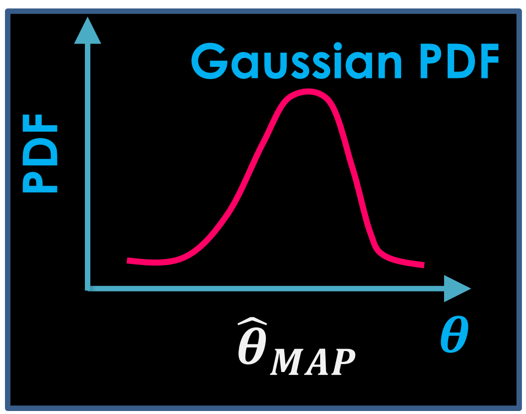


$$P(\theta \mid data) = \frac{P(\theta)\, P(data \mid \theta)}{P(data)}$$

posterior distribution

likelihood of the evidence under any circumstances

probability density function (PDF)



As the amount of data that you collect increases, then the priors plays less and less in terms of determining the posterior

Frequentist vs. Bayesian ≡ MLE vs. MAP

- Maximum Likelihood estimation (MLE)

  Choose value that maximizes the probability of observed data

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} P(D|\theta)$$

- Maximum *a posteriori* (MAP) estimation

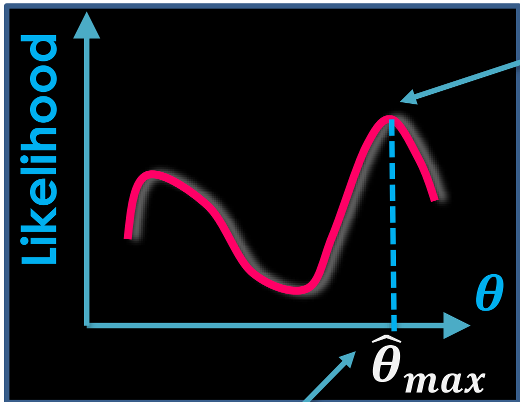
  Choose value that is most probable given observed data and prior belief

$$\begin{aligned}\widehat{\theta}_{MAP} &= \arg\max_{\theta} P(\theta|D) \\ &= \arg\max_{\theta} P(D|\theta)P(\theta)\end{aligned}$$

Parameter Estimation

LIKELIHOOD & LOG-LIKELIHOOD

Likelihood Function, $l\,(\boldsymbol{\theta}|\boldsymbol{data}) \equiv \boldsymbol{P}(\boldsymbol{data}\,|\,\boldsymbol{\theta})$



maximum value of the likelihood function

parameter value that maximizes the likelihood function

Likelihood function:

$$\boxed{l\,(\boldsymbol{\theta}\,|\,\boldsymbol{data}) \equiv \boldsymbol{P}\,(\,\boldsymbol{data}\,|\,\boldsymbol{\theta}\,)}$$

- If data is an **i.i.d.** (independent and identically distributed) sample $\boldsymbol{X} = \{x^t\}, t = 1, \dots, n,$
- Then each instance x^t is drawn from the same distribution (probability density family), defined up to parameters, θ:
  - $x^t \sim p(x, \theta)$
- Hence, due to independence assumption:
  - $l(\theta \mid data) \equiv l(\theta \mid \boldsymbol{X}) \equiv p(\boldsymbol{X} \mid \theta) =$
    $= p(x^1|\theta)\, p(x^2|\theta) \dots p(x^n|\theta)$
    $= \prod_{t=1}^{n} p\,(\,x^t \mid \theta\,)$

A and B are independent:

p(A,B) = p(A)p(B)

Example: Likelihood Function, $l\,(\theta|data)$

Likelihood function: $\boxed{l\,(\theta \mid data) \;\equiv P\,(\,data \mid \theta\,)}$

- Due to independence assumption:
  - $l(\theta \mid data) \equiv l(\,\theta \mid \boldsymbol{X}) \equiv p(\boldsymbol{X} \mid \theta) =$
    $= p(x^1|\theta)\; p(x^2|\theta) \ldots p(x^n|\theta) = \prod_{t=1}^{n} p\,(\,x^t \mid \theta\,)$

- Known: Data, **X**= $\{x^t\}$ = { 5, 10, 7, 4.5, 6.5, 8.7, 9, 6 }; each instance is drawn from the normal (Gaussian) distribution with *unknown* mean and *known* variance $\sigma^2 = 4.0$:
  - $x^t \sim N\,(\mu_X, \sigma^2 = 4.0)$

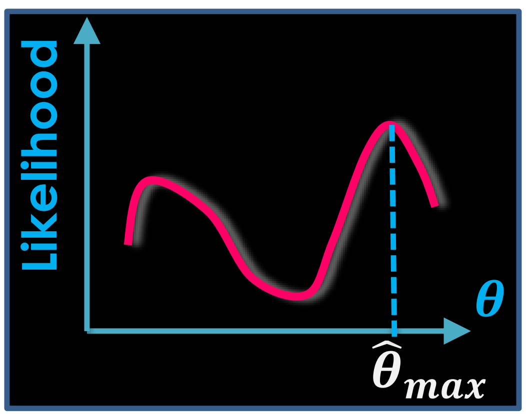
- Unknown Parameter: $\theta = \mu_X$

Which is the largest?

$\boxed{l\,(\theta = 1 \mid X)}$ **= ?**

$\boxed{l\,(\theta = 3 \mid X)}$ **= ?**

$\boxed{l\,(\theta = 7 \mid X)}$ **= ?**



$\boxed{\hat{\theta}_{max}}$ **= ?**

Primer: Gaussian Distribution, $N(\mu, \sigma^2)$

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

Example: Likelihood Function, $l\ (\theta | data)$

- $l(\theta \mid data) \equiv l(\theta \mid \boldsymbol{X}) \equiv p(\boldsymbol{X} \mid \theta) = p(x^1|\theta)\, p(x^2|\theta) \ldots p(x^n|\theta) = \prod_{t=1}^{n} p(x^t \mid \theta)$

- Known: **X**= $\{x^t\}$ = { 5, 10, 7, 4.5, 6.5, 8.7, 9, 6 }:
  - $x^t \sim N(\mu_X, \boldsymbol{\sigma^2 = 4.0})$
- Unknown Parameter: $\theta = \mu_X$

$$\boxed{\boldsymbol{l\ (\theta = 1 \mid X)} = ?}$$

$$\boxed{\boldsymbol{p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}}$$

$$\boldsymbol{l\ (\mu = 1 \mid X)} = \prod_{t=1}^{n} p(x^t \mid \mu = 1, \sigma^2 = 4) =$$

$$= \prod_{t=1}^{n} \boldsymbol{\frac{1}{2\sqrt{2\pi}} e^{-\frac{(x^t-1)^2}{2*4}}} =$$

X= $\{x^t\}$ = { 5, 10, 7, 4.5, 6.5, 8.7, 9, 6 }

$$= \boldsymbol{\frac{1}{2\sqrt{2\pi}} e^{-\frac{(5-1)^2}{2*4}} \times \frac{1}{2\sqrt{2\pi}} e^{-\frac{(10-1)^2}{2*4}} \times \frac{1}{2\sqrt{2\pi}} e^{-\frac{(7-1)^2}{2*4}} \times \cdots \times \frac{1}{2\sqrt{2\pi}} e^{-\frac{(6-1)^2}{2*4}}}$$

What is the issue? – Machine precision! – How to overcome it?

From Likelihood to Log-Likelihood

- We do NOT need to know the value of the likelihood function, $l()$
- We need to have ways to COMPARE $l(\theta|data)$ for different parameter values

$$\boxed{\boldsymbol{l}\,(\boldsymbol{\theta} = \mathbf{1} \mid \boldsymbol{X})} > \boxed{\boldsymbol{l}\,(\boldsymbol{\theta} = \mathbf{3} \mid \boldsymbol{X})} > \boxed{\boldsymbol{l}\,(\boldsymbol{\theta} = \mathbf{7} \mid \boldsymbol{X})}$$

or

$$\boxed{\boldsymbol{l}\,(\boldsymbol{\theta} = \mathbf{1} \mid \boldsymbol{X})} < \boxed{\boldsymbol{l}\,(\boldsymbol{\theta} = \mathbf{3} \mid \boldsymbol{X})} < \boxed{\boldsymbol{l}\,(\boldsymbol{\theta} = \mathbf{7} \mid \boldsymbol{X})}$$

- $l(\theta \mid data) \equiv l(\theta \mid \boldsymbol{X}) \equiv p(\boldsymbol{X} \mid \theta) = p(x^1|\theta)\, p(x^2|\theta) \ldots p(x^n|\theta) = \prod_{t=1}^{n} p\,(x^t \mid \theta)$

If $\boxed{\boldsymbol{l}\,(\boldsymbol{\theta} = \mathbf{1} \mid \boldsymbol{X})} > \boxed{\boldsymbol{l}\,(\boldsymbol{\theta} = \mathbf{3} \mid \boldsymbol{X})} > \boxed{\boldsymbol{l}\,(\boldsymbol{\theta} = \mathbf{7} \mid \boldsymbol{X})}$

then

$$\boxed{\log \boldsymbol{l}\,(\boldsymbol{\theta} = \mathbf{1} \mid \boldsymbol{X})} > \boxed{\log \boldsymbol{l}\,(\boldsymbol{\theta} = \mathbf{3} \mid \boldsymbol{X})} > \boxed{\log \boldsymbol{l}\,(\boldsymbol{\theta} = \mathbf{7} \mid \boldsymbol{X})}$$

Log-Likelihood, $L(\theta|data) \equiv$ log $l(\theta|data)$

Log-Likelihood function:

$$\boxed{L(\theta|data) \equiv log\ l\ (\theta\ |\ data) \ \equiv log\ P\ (\ data\ |\ \theta\ )}$$

- If data is an **i.i.d.** (independent and identically distributed) sample $X = \{x^t\}, t = 1, \ldots, n,$
- Then each instance x^t is drawn from the same distribution (probability density family), defined up to parameters, θ:
  - $x^t \sim p(x, \theta)$

- Hence, due to independence assumption:
  - $L(\theta\ |data\ ) \equiv \log l(\theta\ \ |\ data) \equiv \log l(\ \theta\ |\ X) \equiv \log p(X\ |\ \theta) =$
    $= \log p(x^1|\theta)\ p(x^2|\theta) \ldots p(x^n|\theta)$
    $= \ \sum_{t=1}^{n} \log p\ (\ x^t\ |\ \ \theta\ )$

$$\boxed{L(\theta|data) = \log l(\theta|data) = \sum_{t=1}^{n} \log p\ (\ x^t\ |\ \ \theta\ )}$$

Example: Log-Likelihood, L $(\theta|data)$

- $l(\theta \mid data) \equiv l(\theta \mid \boldsymbol{X}) \equiv p(\boldsymbol{X} \mid \theta) = p(x^1|\theta)\, p(x^2|\theta) \dots p(x^n|\theta) = \prod_{t=1}^{n} p(x^t \mid \theta)$

- Known: **X**= $\{x^t\}$ = { 5, 10, 7, 4.5, 6.5, 8.7, 9, 6 }:
  - $x^t \sim N(\mu_X, \sigma^2 = 4.0)$
- Unknown Parameter: $\theta = \mu_X$

$$\boldsymbol{p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}$$

$\mathbf{L}\,(\boldsymbol{\theta} = \mathbf{1} \mid \boldsymbol{X})$ **= ?**

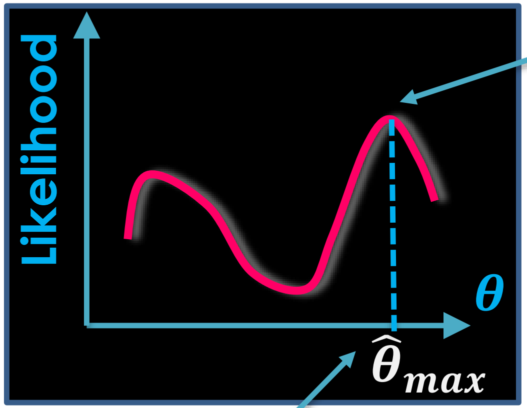
$\mathbf{L}\,(\boldsymbol{\theta} = \mathbf{3} \mid \boldsymbol{X})$ **= ?**

$\mathbf{L}\,(\boldsymbol{\theta} = \mathbf{7} \mid \boldsymbol{X})$ **= ?**

Frequentist Approach

MLE: MAXIMUM LIKELIHOOD ESTIMATION

Maximizing Likelihood Function



maximum value of the likelihood function

parameter value that maximizes the likelihood function

$$\max_{\theta}(Likelihood\_Function)$$

equivalent to

$$\max_{\theta} P\ (data\ |\theta\ )$$

argmax(): returns the value of the argument / parameter, for which the likelihood function attains its maximum

$$\hat{\theta}_{max} = \operatorname*{argmax}_{\theta} P(data\ |\theta)$$

equivalent to

$$\hat{\theta}_{MLE} = \operatorname*{argmax}_{\theta} P(data\ |\theta)$$

maximum likelihood estimator **for the parameter θ**