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

OAK RIDGE National Laboratory

NC STATE UNIVERSITY Department of Computer Science

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
- Bayesian Parameter Estimation
  - MAP vs. Full Bayesian Treatment
  - Prediction with Bayesian Parameter Estimators
  - Bayesian Regression vs. MLE 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**

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

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**

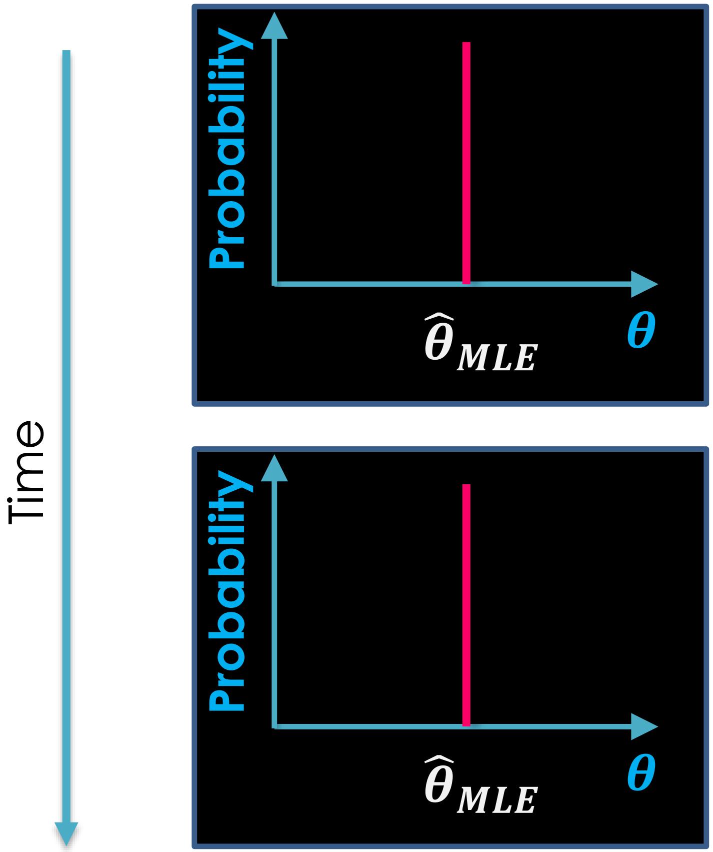
Informally, Frequentist vs. Bayesian

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.

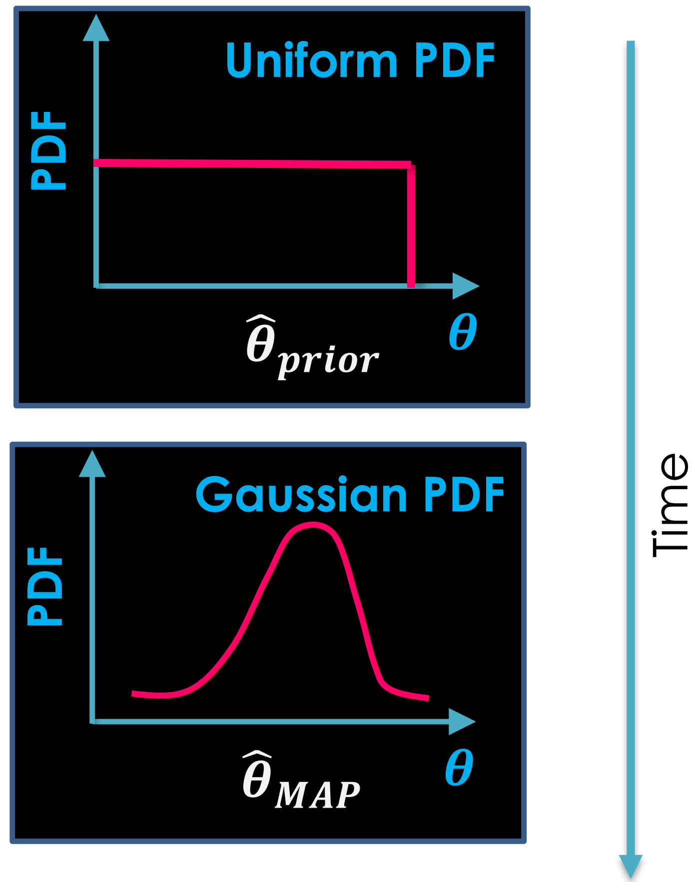
Visually, Frequentist vs. Bayesian

Frequentist's View Point



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

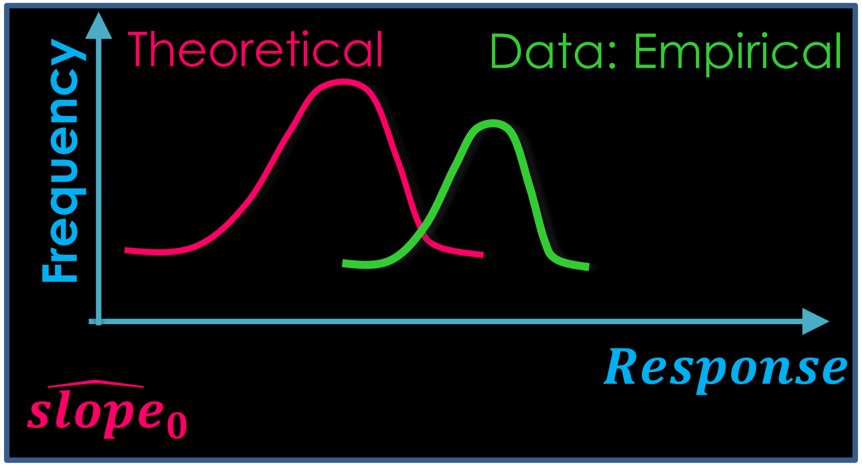
Bayesian View Point

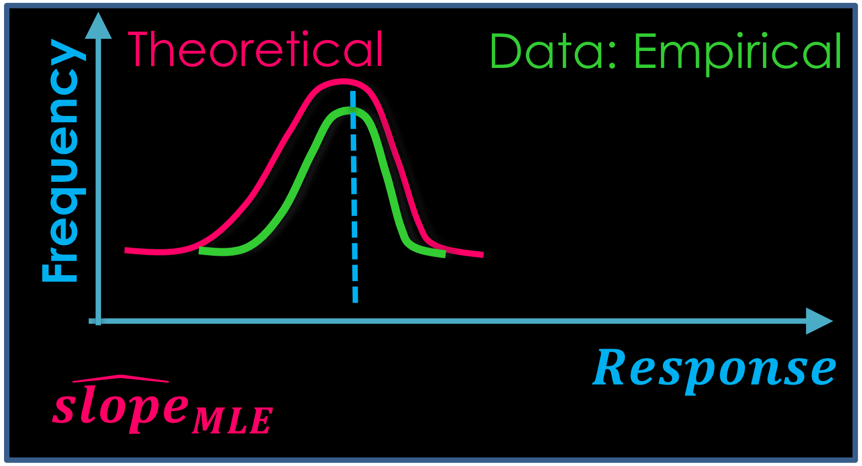


- Parameter, θ is a **random variable** with a probability distribution

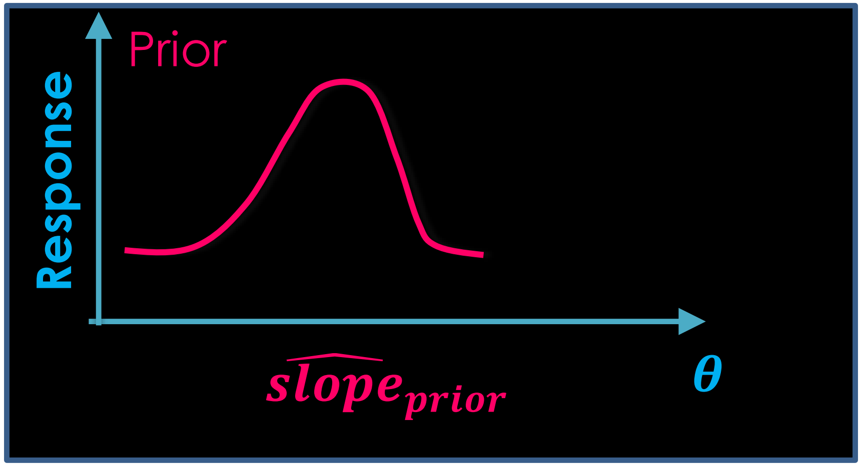
Example: MLE

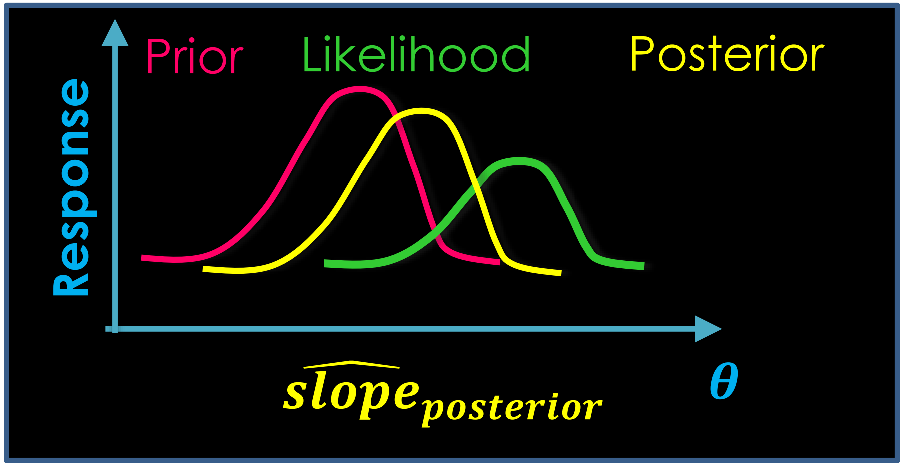
- **MLE Example**: In Linear Regression
  - We estimate the most likely value for the slope and intercept parameters
    - **how well** $slope_{MLE}$ and $intercept_{MLE}$ **fit** the given **data**
  - Make a single prediction for the most likely response value as specified by ($slope_{MLE}$ and $intercept_{MLE}$)





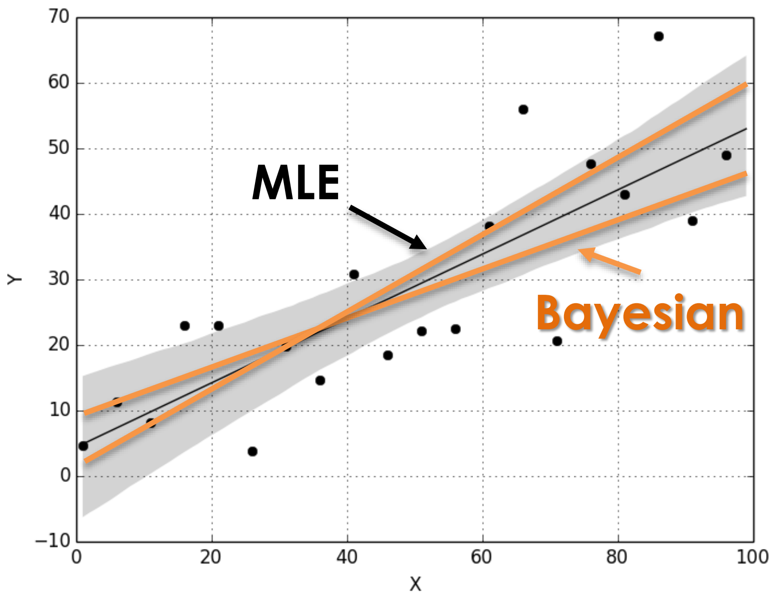
Example: Bayesian





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

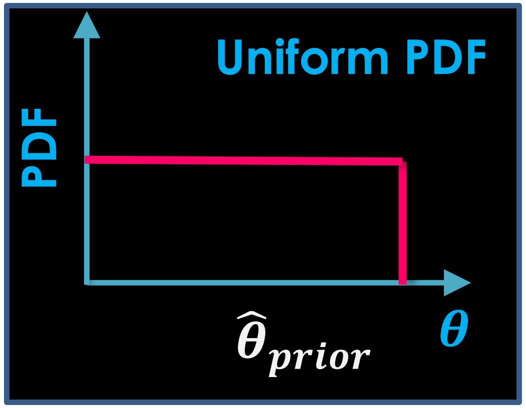
- **Prediction:** Instead of using a single θ estimate in prediction, a set of possible θ values is generated as defined by the posterior
  - To use all of them in prediction,
  - Weighted by how likely each of the value is (i.e., sum or integrate)

- **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

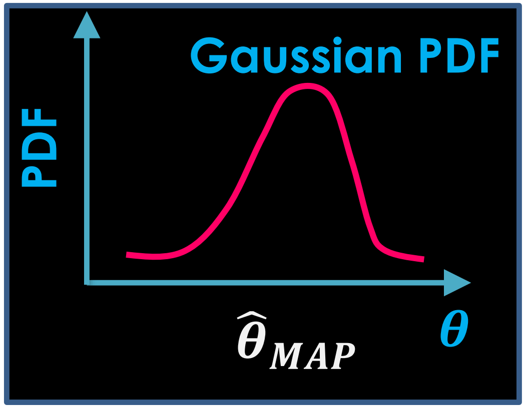


$$P(\theta \mid data) = \frac{P(\theta)\, P(data \mid \theta)}{P(data)}$$

posterior distribution

likelihood of the evidence under any circumstances

probability density function (PDF)

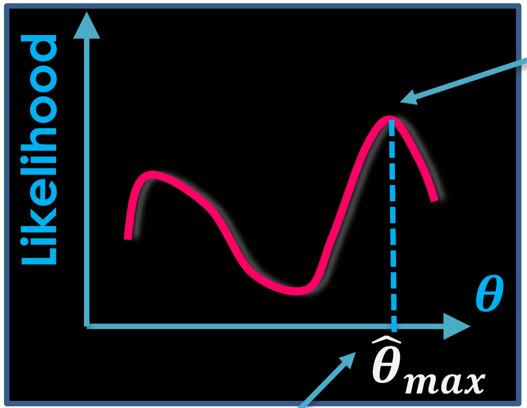


As the amount of data that you collect increases, then the priors plays less and less in terms of determining the posterior

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

A and B are independent:

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

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

Likelihood function: $\boxed{l\ (\theta\ |\ data)\ \equiv P\ (\ data\ |\ \theta\ )}$

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

- <u>Known</u>: 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 $\sigma^2 = 4.0$:
  - $x^t \sim N\ (\mu_X, \sigma^2 = 4.0)$

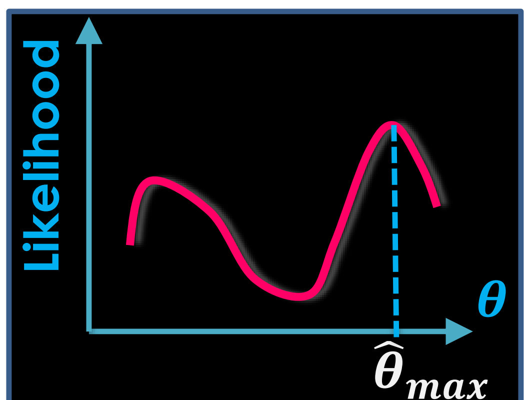
- <u>Unknown Parameter</u>: $\theta = \mu_X$

Which is the largest?

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

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

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



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

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

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

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

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

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

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

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

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

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

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

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

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

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

or

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

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

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

then

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

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

Log-Likelihood function:

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

- If data is an **i.i.d.** (independent and identically distributed) sample $\boldsymbol{X} = \{x^t\}, t = 1, \dots, n,$
- 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:
  - $L(\theta\ |data\ ) \equiv \log l(\theta\ \ |\ data) \equiv \log l(\ \theta\ |\ \boldsymbol{X}) \equiv \log p(\boldsymbol{X}\ |\ \theta) =$
    $= \log p(x^1|\theta)\ p(x^2|\theta) \dots p(x^n|\theta)$
    $=\ \sum_{t=1}^{n} \log p\ (\ x^t\ |\ \ \theta\ )$

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

Log-Likelihood $L(\theta|data)$ for Gaussian Density

Log-Likelihood function:

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

If data is an **i.i.d.** sample $X = \{x^t\}, t = 1, \ldots, n,$

Gaussian (Normal) Density:

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

$$L(\mu, \sigma^2 | X) = -\frac{n}{2} log(2\pi) - n log(\sigma) - \frac{1}{2\sigma^2} \sum_{t=1}^{n} (x^t - \mu)^2$$

log() is natural log

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: $\theta = \mu_X$

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

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

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

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

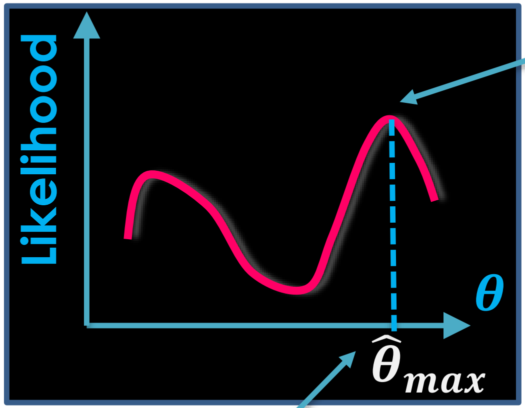
Frequentist Approach

MLE: MAXIMUM LIKELIHOOD ESTIMATION

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{MLE}} = \underset{\boldsymbol{\theta}}{\mathbf{arg\,max}}\, \boldsymbol{L}(\boldsymbol{\theta}|\boldsymbol{data}) = \underset{\boldsymbol{\theta}}{\mathbf{arg\,max}}\, \mathbf{log}\, \boldsymbol{P}(\boldsymbol{data}|\boldsymbol{\theta})$$

Choose parameter estimator that maximizes the ***likelihood*** *of observed data, or maximizes the fit of the observed data to the theoretical PDF for this parameter value.*

Maximizing (Log-)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

Primer: Derivative Formulas

Derivative of a constant	$\frac{dc}{dx} = 0$
Derivative of constant multiple	$\frac{d}{dx}(cu) = c\,\frac{du}{dx}$
Derivative of sum or difference	$\frac{d}{dx}(u \pm v) = \frac{du}{dx} \pm \frac{dv}{dx}$
Product Rule	$\frac{d}{dx}(uv) = u\,\frac{dv}{dx} + v\,\frac{du}{dx}$
Quotient Rule	$\frac{d}{dx}\left(\frac{u}{v}\right) = \frac{v\,\frac{du}{dx} - u\,\frac{dv}{dx}}{v^2}$
Chain Rule	$\frac{dy}{dx} = \frac{dy}{du}\,\frac{du}{dx}$

Primer: Derivative Formulas (cont.)

u = f(x): a function of **x**. ***a*** is a constant; **n** is a integer.

$$\frac{d}{dx}\, x^n = nx^{n-1} \qquad \frac{d}{dx}\, u^n = nu^{n-1}\, \frac{du}{dx}$$

$$\frac{d}{dx}\, a^x = (\ln a)\, a^x \qquad \frac{d}{dx}\, a^u = (\ln a)\, a^u\, \frac{du}{dx}$$

$$\frac{d}{dx}\, e^x = e^x \qquad \frac{d}{dx}\, e^u = e^u\, \frac{du}{dx}$$

$$\frac{d}{dx} \log_a x = \frac{1}{(\ln a)\, x} \qquad \frac{d}{dx} \log_a u = \frac{1}{(\ln a)\, u}\, \frac{du}{dx}$$

$$\frac{d}{dx} \ln x = \frac{1}{x} \qquad \frac{d}{dx} \ln u = \frac{1}{u}\, \frac{du}{dx}$$

Log-Likelihood $L(\theta|data)$ for Gaussian Density

Log-Likelihood function:

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

If data is an **i.i.d.** sample $\boldsymbol{X} = \{x^t\}, t = 1, \dots, n,$

Gaussian (Normal) Density:

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

$$\boldsymbol{L(\mu, \sigma^2 | X) = -\frac{n}{2} log(2\pi) - n log(\sigma) - \frac{1}{2\sigma^2} \sum_{t=1}^{n} (x^t - \mu)^2}$$

log() is natural log

MLE Parameter Estimation for Gaussian Density

Log-Likelihood function for Gaussian Density:

$$L(\mu, \sigma^2 | X) = -\frac{n}{2} log(2\pi) - n log(\sigma) - \frac{1}{2\sigma^2} \sum_{t=1}^{n} (x^t - \mu)^2$$

log() is natural log

MLE estimation of $\theta = (\mu, \sigma)$:

$$0 = \frac{\partial L(\mu, \sigma^2 | X)}{\partial \mu} = \frac{\partial}{\partial \mu} \sum_{t=1}^{n} (x^t - \mu)^2 = \sum_{t=1}^{n} (x^t - \mu) = \sum_{t=1}^{n} x^t - n\mu$$

MLE estimator of the parameter μ of the Gaussian distribution is the sample mean.

$$\boxed{\hat{\mu}_{MLE} = \frac{1}{n} \sum_{t=1}^{n} x^t = \bar{X}}$$

MLE Parameter Estimation for Gaussian Density

Log-Likelihood function for Gaussian Density:

$$L\left(\mu, \sigma^2 \middle| X\right) = -\frac{n}{2} log(2\pi) - nlog(\sigma) - \frac{1}{2\sigma^2}\sum_{t=1}^{n}(x^t - \mu)^2$$

log() is natural log

MLE estimation of $\theta = (\mu, \sigma)$:

$$0 = \frac{\partial L\left(\mu, \sigma^2 \middle| X\right)}{\partial \sigma} = \frac{\partial}{\partial \sigma}\left[-nlog(\sigma) - \frac{1}{2\sigma^2}\sum_{t=1}^{n}(x^t - \mu)^2\right] = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{t=1}^{n}(x^t - \mu)^2$$

MLE estimator of the parameter σ^2 of the Gaussian distribution:

$$\widehat{\sigma^2}_{MLE} = \frac{1}{n}\sum_{t=1}^{n}(x^t - \mu)^2$$

Prediction with the MLE Estimators for Gaussian

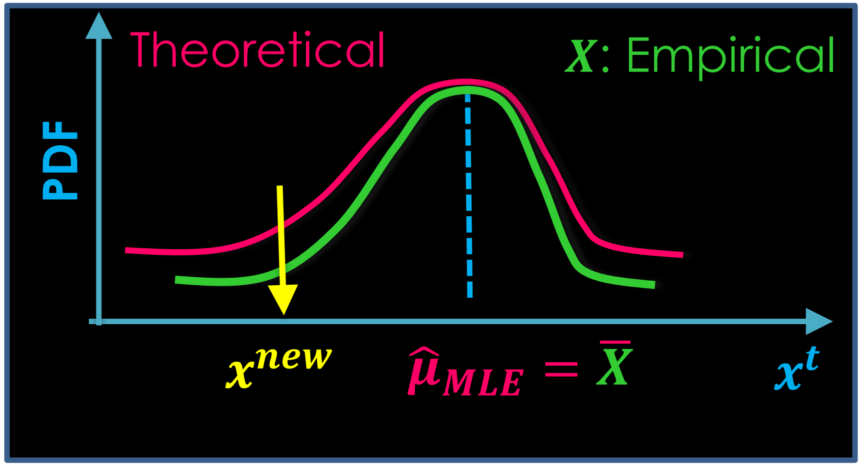
MLE Estimators for Gaussian Density $\widehat{\theta}_{MLE} = (\hat{\mu}_{MLE}, \widehat{\sigma^2}_{MLE})$:

$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_{t=1}^{n} x^t = \bar{X}$$

$$\widehat{\sigma^2}_{MLE} = \frac{1}{n}\sum_{t=1}^{n}(x^t - \mu)^2$$

Prediction of the probability that x^{new} comes from the same Gaussian distribution as the training i.i.d. sample $X = \{x^t\}, t = 1, \dots, n$:

$$p(x^{new} \mid X) = p(x^{new} \mid \widehat{\theta}_{MLE})$$



$$p(x^{new} \mid X, \mu_{MLE}, \sigma^2_{MLE}) = \frac{1}{\sqrt{2\pi\sigma^2_{MLE}}} e^{-\frac{(x^{new} - \mu_{MLE})^2}{2\sigma^2_{MLE}}}$$

Bayesian Approach

MAP: MAXIMUM A POSTERIOR ESTIMATION

$$\hat{\boldsymbol{\theta}}_{MAP} = \arg\max_{\boldsymbol{\theta}} P(\boldsymbol{\theta} | \boldsymbol{data})$$
$$= \arg\max_{\boldsymbol{\theta}} P(\boldsymbol{\theta}) P(\boldsymbol{data} | \boldsymbol{\theta})$$

Choose parameter estimator that maximizes the ***posterior*** *probability given observed data and prior belief.*

Bayesian Inference: Posterior Density

- The ***prior density***, $p(\theta)$, tells us the likely values that θ may take before looking at the sample.

- The ***likelihood density***, $p(X \mid \theta)$, tells us the likely values that θ may take by looking at the sample, i.e., how likely the sample X is if the parameter of the distribution takes the value of θ.

- Thus, the Bayes' rule, the ***posterior density***, $p(\theta \mid X)$, tells us the likely θ values after looking at the sample and taking priors into account:

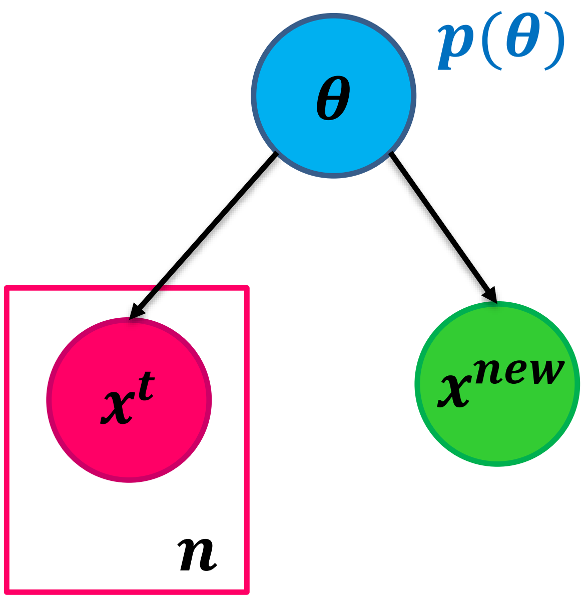
$$\boxed{\boldsymbol{p(\theta \mid X) = \frac{p(X \mid \theta)\, p(\theta)}{p(X)} = \frac{p(X \mid \theta)\, p(\theta)}{\int p(X \mid \theta')\, p(\theta')d\theta'}}}$$

- Given the ***posterior density***, $p(\theta \mid X)$, for the training data $\boldsymbol{X}$, the ***prediction*** of the probability of a new observation, x^{new}, to come from the same distribution:

Bayesian Inference: Generative Model

- **Generative Model:** Represents **how the data is generated**:
  - First, sample θ from $p(\theta)$
  - Then generate the training instances x^t by sampling from $p(x \mid \theta)$
  - Finally, generate the new instance x^{new}

$$\boxed{p(x^{new}, X, \theta) = p(\theta)p(X \mid \theta)p(x^{new}|\theta)} \quad **$$



$$X = \{x^t\}_{t=1}^{n}$$

To estimate probability for the x^{new} given the training sample X:

$$p(x^{new} \mid X) = \frac{p(x^{new}, X)}{p(X)} =$$

$$= \frac{\int p(x^{new}, X, \theta) d\theta}{p(X)} =$$

$$\overset{**}{=} \frac{\int p(\theta)p(X \mid \theta)p(x^{new}|\theta)\, d\theta}{p(X)} =$$

$$= \int p\,(\theta \mid X\,)\,p\,(\,x^{new} \mid\, \theta)\, d\theta$$

Bayesian Inference: Prediction

$$\boxed{p(\boldsymbol{\theta} \mid \boldsymbol{X}) = \frac{p(\boldsymbol{X} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\boldsymbol{X})}}$$

- **Prediction using Generative Model:** Given the ***posterior density***, $p(\theta \mid X)$, derived from the training sample $\boldsymbol{X}$ and the priors, the probability of a new observation, x^{new}, to come from the same distribution:

The estimate for the probability of x^{new} given $\boldsymbol{X}$ as **the weighted sum** of estimates using all possible values of θ weighted by how likely each θ is, given the sample $\boldsymbol{X}$.

$$p(\boldsymbol{x}^{new} \mid \boldsymbol{X}) = \int p(\boldsymbol{\theta} \mid \boldsymbol{X})\, p(\boldsymbol{x}^{new} \mid \boldsymbol{\theta})\, d\boldsymbol{\theta}$$

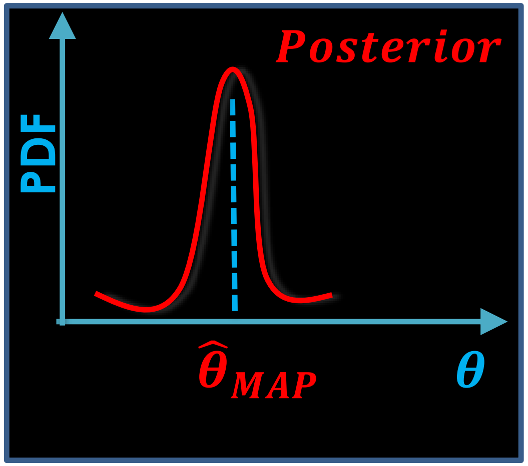
if θ is discrete valued:

$$p(\boldsymbol{x}^{new} \mid \boldsymbol{X}) = \sum_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid \boldsymbol{X})\, p(\boldsymbol{x}^{new} \mid \boldsymbol{\theta})$$

What if the posterior is NOT easy to integrate?

$$p(x^{new} \mid X) = \int p(\theta \mid X)\, p(x^{new} \mid \theta)\, d\theta$$

Assumption: The **posterior** makes a very **narrow peak** around a single point



- Then use the **mode** of the posterior:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta \mid X) = \arg\max_{\theta} P(\theta)P(X \mid \theta)$$

- To make the **prediction**:

$$p_{MAP}(x^{new} \mid X) = p(x^{new} \mid \hat{\theta}_{MAP})$$

Frequentist vs. Bayesian ≡ MLE vs. MAP

MLE: Maximum (Log-)Likelihood Estimation:

$$\widehat{\boldsymbol{\theta}}_{MLE} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\boldsymbol{data}) = \arg\max_{\boldsymbol{\theta}} \log P(\boldsymbol{data}|\boldsymbol{\theta})$$

*Choose parameter estimator that maximizes the **likelihood** of observed data, or maximizes the fit of the observed data to the theoretical PDF for this parameter value.*

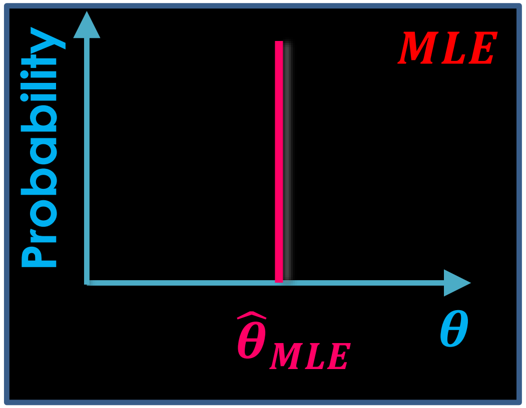
MAP: Maximum A Posterior Estimation:

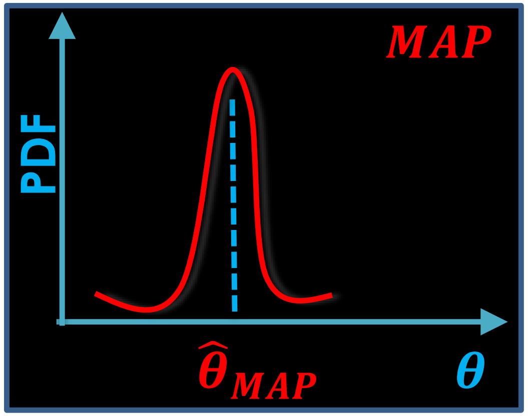
$$\widehat{\boldsymbol{\theta}}_{MAP} = \arg\max_{\boldsymbol{\theta}} P(\boldsymbol{\theta}|\boldsymbol{data}) = \arg\max_{\boldsymbol{\theta}} P(\boldsymbol{\theta})P(\boldsymbol{data}|\boldsymbol{\theta})$$

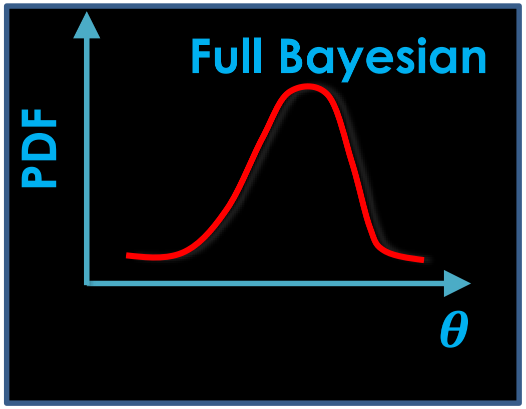
*Choose parameter estimator that maximizes the **posterior** probability given observed data and prior belief.*

Summary: Parameter Estimation & Prediction

Assumptions	Parameter Estimation	Prediction, $p(x^{new} \| X)$
MLE	$\hat{\theta}_{MLE} = \arg\max_{\theta} p(X \mid \theta)$	$p_{MLE}(x^{new} \mid X) = p(x^{new} \mid \hat{\theta}_{MLE})$
Bayesian: MAP (narrow peak)	$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta \mid X)$	$p_{MAP}(x^{new} \mid X) = p(x^{new} \mid \hat{\theta}_{MAP})$
Bayesian: Full Treatment	$p(\theta \mid X) = \dfrac{p(X \mid \theta)\, p(\theta)}{p(X)}$	$p(x^{new} \mid X) = \sum_{\theta} p(\theta \mid X)\, p(x^{new} \mid \theta)$

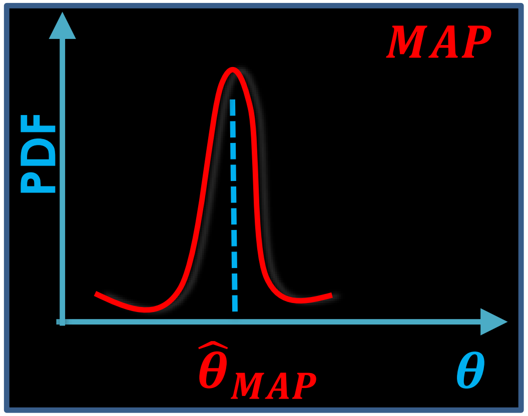


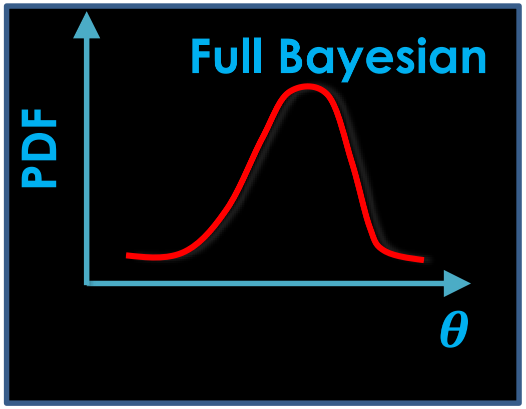




Bayesian Estimates & Predictions

Assumptions	Parameter Estimation	Prediction, $p(x^{new}\|X)$
Bayesian: Approximate the Integral	????	????
Bayesian: MAP (narrow peak)	$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta \mid X)$	$p_{MAP}(x^{new} \mid X) = p(x^{new} \| \hat{\theta}_{MAP})$
Bayesian: Full Integral, if possible to $\int$	$p(\theta \mid X) = \dfrac{p(X \mid \theta)\, p(\theta)}{p(X)}$	$p(x^{new} \mid X) = \sum_{\theta} p(\theta \mid X)\, p(x^{new} \mid \theta)$





Linear Regression

MLE APPROACH

Simple Linear Regression Model

- **Response** variable (vector): $\vec{\boldsymbol{r}} = \{r^t\}_{t=1}^{n}, r^t \in \mathbb{R}$
- **Predictor** variables (matrix): $\boldsymbol{X_{n\times d}} = \{x^t\}_{t=1}^{n}, x^t = (\mathbf{1}, x_2^t, x_3^t, \dots, x_d^t) \in \mathbb{R}^d$
- **Data:** $\chi = [\,\boldsymbol{X_{n\times d}}, \vec{\boldsymbol{r}}\,]$

- Regression coefficients/weights (vector): $\vec{w} = \{w_k\}_{k=1}^{d}, w_k \in \mathbb{R}$
- Known: Precision of the additive noise (random variable): $\epsilon \sim N(0, \frac{1}{\gamma})$

$$\boldsymbol{r^t = \vec{w}^T x^t + \epsilon}$$

$$\boldsymbol{p(r^t | x^t, \vec{w}, \gamma) \sim N\left(\vec{w}^T x^t, \ \frac{1}{\gamma}\right)}$$

(Log-)Likelihood for Linear Regression (LR)

$$L(\overrightarrow{w} \mid \chi) \equiv log\, p(\chi \mid \overrightarrow{w}) = log\, p(\vec{r}, X \mid \overrightarrow{w}) = log\, p(\vec{r} \mid X, \overrightarrow{w}\,) + log\, p(X)$$

$$log\, p(\vec{r} \mid X, \overrightarrow{w},\ \ \gamma\,) = log\, \prod_t p(r^t \mid x^t, \overrightarrow{w}, \gamma)$$

$$p(r^t \mid x^t, \overrightarrow{w}, \gamma) \sim N\, (\overrightarrow{w}^T x^t,\ \frac{1}{\gamma}\,)$$

$$log\, p(\vec{r} \mid X, \overrightarrow{w},\ \ \gamma\,) = -n log\left(\sqrt{2\pi}\right) + n\, log\, \sqrt{\gamma}\ - \frac{\gamma}{2} \sum_t \left(r^t - w^T x^t\right)^2$$

$$Error = \sum_t \left(r^t - \overrightarrow{w}^T x^t\right)^2 = (\vec{r}\ - X\overrightarrow{w})^T(\vec{r}\ - X\overrightarrow{w})$$

$$Error = \vec{r}^T \vec{r}\ - 2\overrightarrow{w}^T X^T \vec{r} + \overrightarrow{w}^T \overrightarrow{(X^T X) w}$$

MLE: Maximizing (Log-)Likelihood for LR

$$\hat{w}_{MLE} = \arg\max_{\theta} p(\chi | \vec{w})$$

$$\Updownarrow$$

equivalent to

$$\hat{w}_{MLE} = \arg\min_{\theta} Error(\vec{w})$$

$$\mathbf{0} = \frac{\partial}{\partial w} Error(\vec{w}) = \frac{\partial}{\partial w} \sum_t \left(r^t - \vec{w}^T x^t\right)^2 = \frac{\partial}{\partial w} (\vec{r} - X\vec{w})^T (\vec{r} - X\vec{w})$$

$$0 = \frac{\partial}{\partial w} Error(\vec{w}) = \frac{\partial}{\partial w} (\vec{r}^T \vec{r} - 2\vec{w}^T X^T \vec{r} + \vec{w}^T \overrightarrow{(X^T X) w})$$

$$0 = -2X^T \vec{r} + 2\,\overrightarrow{(X^T X) w})$$

$$\boxed{\hat{w}_{MLE} = \left(X^T X\right)^{-1} X^T \vec{r}}$$

Prediction with MLE-based LR model

$$\boxed{\hat{w}_{MLE} = \left(X^T X\right)^{-1} X^T \vec{r}}$$

LR model: $r^t = \vec{w}^T x^t + \epsilon$

$$\boxed{r^{new} = \hat{w}_{MLE}^T x^{new}}$$

Modeling under Uncertainty

BAYESIAN REGRESSION

Posterior Gaussian Density for LR Parameters

- **Conjugate Prior** for the parameters of LR:

$$p(\vec{w}) \sim N\left(0, \frac{1}{\alpha} I_{d\times d}\right)$$

  - we expect parameters to be close to 0 with spread inversely proportional to α
  - when $\alpha \to 0$, then we have a flat prior and $\hat{w}_{MAP}$ converges to $\hat{w}_{MLE}$

- **Posterior** for the parameters of LR for the training i.i.d. sample of size n:

$$\boxed{p(\vec{w} \mid X, \vec{r}) \sim N(\mu_n, \Sigma_n)}$$

$$\mu_n = \gamma \Sigma_n X^T \vec{r}$$

$$\Sigma_n = \left(\alpha I + \gamma X^T X\right)^{-1}$$

- **Prediction** using full posterior integration:

$$\boxed{r^{new} = \int \left(\vec{w}^T x^{new}\right) p(\vec{w} \mid X, \vec{r}) dw}$$

MAP LR Estimator for the Posterior Gaussian

- **Posterior** for the parameters of LR for the training i.i.d. sample of size n:

$$\boxed{p(\vec{w} \mid X, \vec{r}) \sim N\,(\mu_n, \Sigma_n)}$$

$$\mu_n = \gamma \Sigma_n X^T \vec{r}$$

$$\Sigma_n = \left(\alpha I + \gamma X^T X\right)^{-1}$$

A point estimate: $$\boxed{\hat{w}_{MAP} = \mu_n = \gamma\left(\alpha I + \gamma X^T X\right)^{-1} X^T \vec{r}}$$

$$\boxed{r^{new} = \hat{w}_{MAP}^T x^{new}}$$

Replace the posterior density with a single point

$$\log p(\vec{w} \mid X, \vec{r}) \sim \log p(\vec{r} \mid \vec{w}, X) + \log p(w)$$

$$\sim -\frac{\gamma}{2} \sum_t \left(r^t - \vec{w}^T x^t\right)^2 - \frac{\alpha}{2} \vec{w}^T \vec{w}$$

$$0 = \frac{\partial}{\partial w} \log p(\vec{w} \mid X, \vec{r}) \rightarrow \hat{w}_{MAP} = \gamma\left(\alpha I + \gamma X^T X\right)^{-1} X^T \vec{r}$$

Linear Regression: MLE vs. Bayesian Approach

$$\hat{w}_{MLE} = \left(X^T X\right)^{-1} X^T \vec{r}$$

$$r^{new} = \hat{w}_{MLE}^T x^{new}$$

$$\hat{w}_{MAP} = \mu_n = \gamma\left(\alpha I + \gamma X^T X\right)^{-1} X^T \vec{r}$$

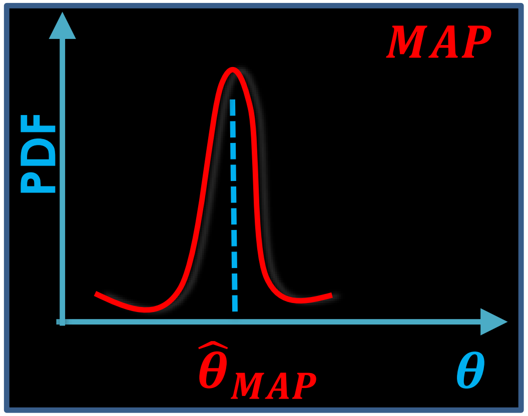
$$r^{new} = \hat{w}_{MAP}^T x^{new}$$

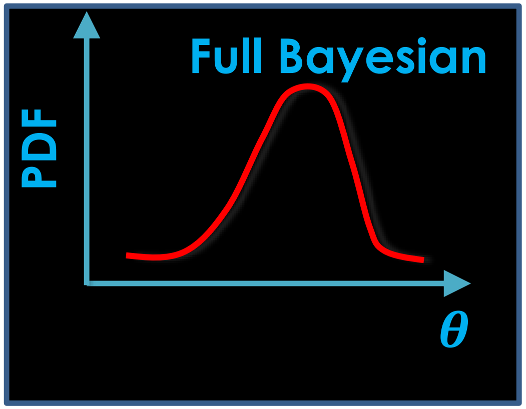
$$p(\vec{w} \mid X, \vec{r}) \sim N\,(\mu_n, \Sigma_n)$$

$$r^{new} = \int \left(\vec{w}^T x^{new}\right) p(\vec{w} \mid X, \vec{r})\, dw$$

What if both integration & MAP are not possible?

Assumptions	Parameter Estimation	Prediction, $p(x^{new}\|X)$
Bayesian: Approximate the Integral	????	????
Bayesian: MAP (narrow peak)	$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta \mid X)$	$p_{MAP}(x^{new} \mid X) = p(x^{new}\|\hat{\theta}_{MAP})$
Bayesian: Full Integral, if possible to $\int$	$p(\theta \mid X) = \frac{p(X \mid \theta)\, p(\theta)}{p(X)}$	$p(x^{new} \mid X) = \sum_{\theta} p(\theta \mid X)\, p(x^{new} \mid \theta)$

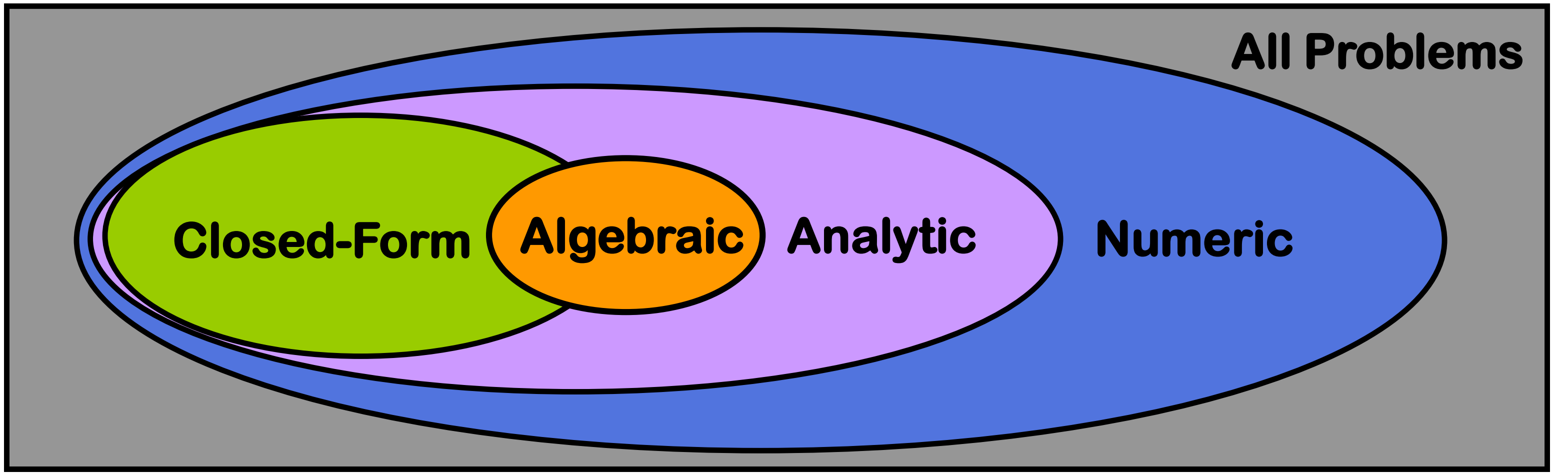




Optimization Problems

CLOSED-FORM, ALGEBRAIC, ANALYTIC, NUMERIC SOLUTIONS

Classes of Problems



Closed-Form Expression

- **A closed-form mathematical expression:**
  - Evaluated in a finite number of operations.
  - Expressed:
    - in terms of constants, variables, "well-known" operations (e.g., + − × ÷), and functions (e.g., n^{th} root, logs, exp, trigonometric functions, and inverse hyperbolic functions
    - but **NOT** in terms of **limits, integrals, infinite series**

- **Tractable Problems**:
  - Can be solved in terms of a closed-form expression
  - Example: $ax^2 + bx + c = 0$ is a tractable problem; its solution is in a closed-form

- **CDF:** Many cumulative distribution functions (CDF) can **NOT** be expressed in closed-form:
  - Ways around this issue: To consider special functions such as the error function or gamma function.

src: Wiki

Analytic Mathematical Expression

- **An analytic expression:**
  - Constructed using well-known operations that lend themselves readily to calculation
  - Expressed:
    - in terms of constants, variables, "well-known" operations (e.g., + − × ÷), and functions (e.g., n^{th} root, logs, exp, trigonometric functions, and inverse hyperbolic functions,
    - may include **infinite series**, **Gamma and Bessel functions**,
    - but **NOT** **limits, integrals**.
- **Tractable Problems**:
  - Can be solved in terms of a closed-form expression
  - Example: $ax^2 + bx + c = 0$ is a tractable problem; its solution is in a closed-form

- **CDF:** Many cumulative distribution functions (CDF) can **NOT** be expressed in closed-form:
  - Way around this issue: Consider special functions such as the error function or gamma function.

src: Wiki

Algebraic Expression

- **An algebraic expression is an analytic expression:**
  - Expressed only in terms of the algebraic operations (addition, subtraction, multiplication, division and exponentiation to a rational exponent) and rational constants

src: Wiki

Numeric Algorithms

- **Numeric algorithms use numeric approximations:**
  - ***Discretization*** for numeric integration
  - ***Numerical differentiation***
  - ***Iterative*** methods (e.g., Newton's method) for optimization
  - **Numerical interpolation, extrapolation, smoothing**

src: Wiki