Statistical Problem: A Generic Model

Data Modelling:

Data (Observations) Set:	$\mathbb{X}_n := \{X_1, \ldots, X_n \quad \mid \quad X_i \in \mathcal{X} \subset \mathbb{R}^k\}$
Observation Space	$:= \mathcal{X}$
Assume Data IID:	$X_1, \ldots, X_n \quad \sim_{iid}$ copies of "generic" $\quad X$
Distribution of X:	$X \sim P \quad P \in \mathcal{F}$
Family of Distribution	$:= \mathcal{F}$
Distribution of $\mathbb{X}_n$:	$\mathbb{X}_n \sim \mathbb{P} := \underbrace{P \times \cdots \times P}_{\text{n -times}}$

Remark

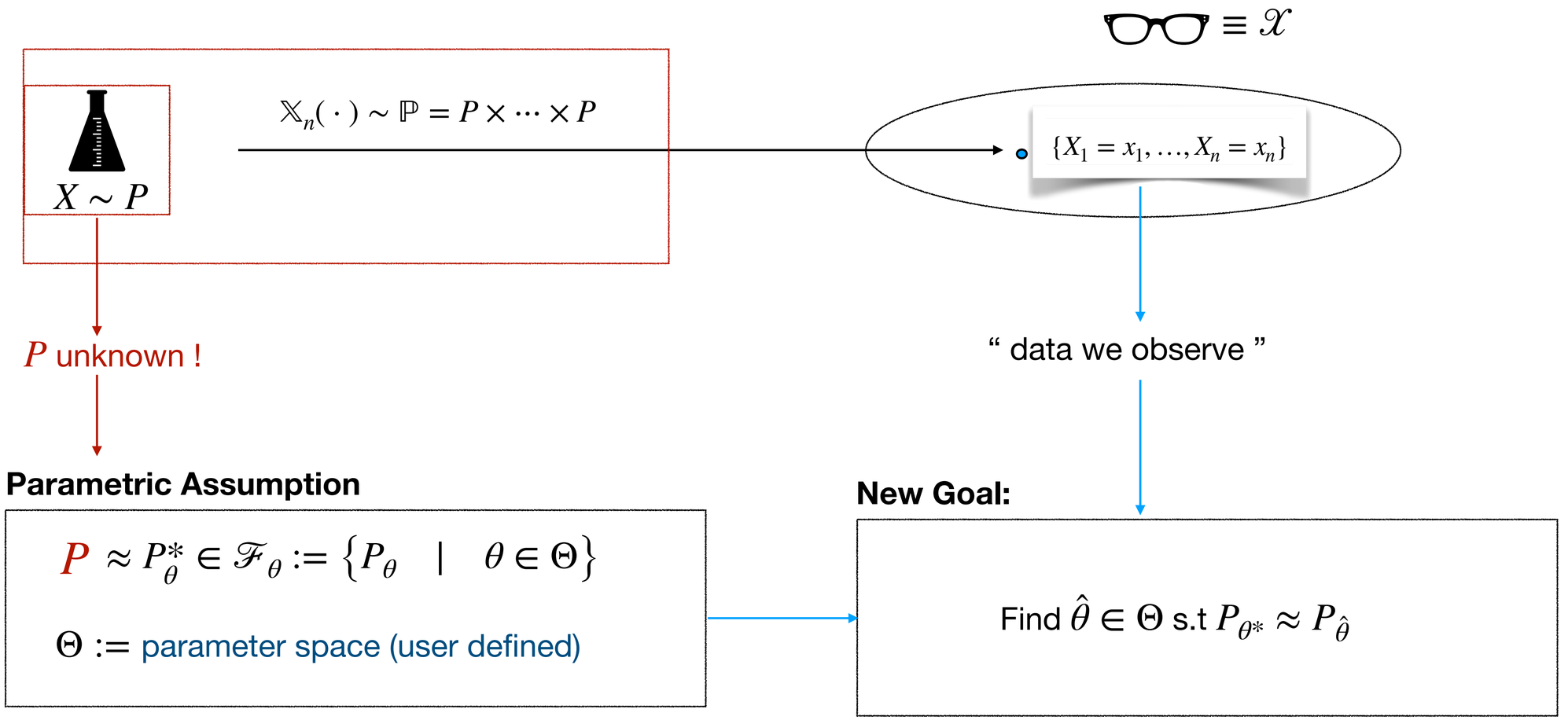
Knowing $P \equiv$ Knowing the problem

Goal:

FIND P

ETH zürich

Statistical Problem: A Parametric Model



Statistical Problem: A Parametric Model : Tossing Example

Given a set of n outcomes from tossing a coin n time, is the coin fair ?

$\mathbb{D}_{coin} := \{\text{head}, \ldots, \text{tail}\}$

"Tossing a coin "

$=: X \in \mathcal{X} = \{\text{head}, \text{tail}\} \equiv \{0,1\}$

Parametric Assumption:

$X \sim Ber(\theta) =: \mathcal{F}_\theta \quad \theta \in \Theta := [0,1]$

Family of densities:

$\mathcal{P}_\theta = \{p_\theta(X = x) = \theta^x(1-\theta)^{1-x} \quad , \quad x \in \{0 = \text{head}, 1 = \text{tail}\}\}$

Data Set as R.V

$\mathbb{X}_n = \{\underbrace{X_1, \ldots, X_n}_{iid \sim X}\} \quad X_i := i^{th}$ time the coin is tossed

Distribution of dataset (Densities Family)

$\mathcal{D}_\theta = \{\mathbb{P}_\theta = \prod_{i \in [n]} P_\theta\} \implies p_\theta(\mathbb{X}_n = \mathbb{D}_{coin}) = \prod_{i \in [n]} \theta^{x_i}(1-\theta)^{1-x_i}$

$\mathbb{X}_n \quad \underset{\text{realisation}}{=} \quad \mathbb{D}_{coin}$

Estimate θ from $\mathbb{D}_n$

$\theta \approx_? 0.5$

Statistical Problem: **Likelihood Function**

ASSUMPTIONS Slide 1 Data Modelling:

Words:

Data Set IID realised

Distribution Parametrised

Parameter Space Θ user defined

Density well defined

Maths:

$$\mathbb{X}_n = \mathbb{D}_n = \{X_1 = x_1, \ldots, X_n = x_n\}$$

$$\mathbb{X}_n \sim \mathbb{P}_\theta := P_\theta \times \cdots \times P_\theta \quad \theta \in \Theta$$

$$p_\theta(\mathbb{D}_n) = \prod_{i \in [n]} p_\theta(x_i)$$

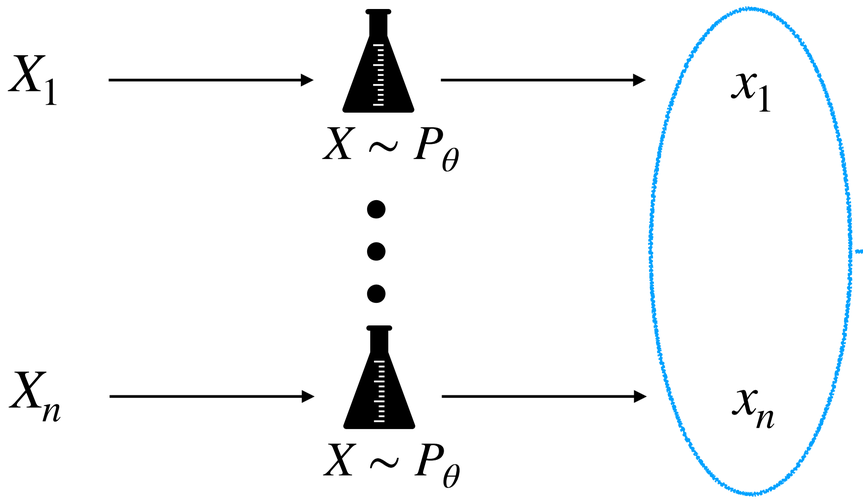
Definition: Likelihood Function of data set (realised) $\mathbb{D}_n$:

$$L_{\mathbb{D}_n} : \Theta \to \mathbb{R}$$

$$L_{\mathbb{D}_n}(\theta) := \prod_{i \in [n]} p_\theta(x_i)$$

ETH zürich

Statistical Problem: **Likelihood Function interpretation**



$\mathbb{D}_n$

$$L_{\mathbb{D}_n}(\theta) := \prod_{i \in [n]} p_\theta(x_i)$$

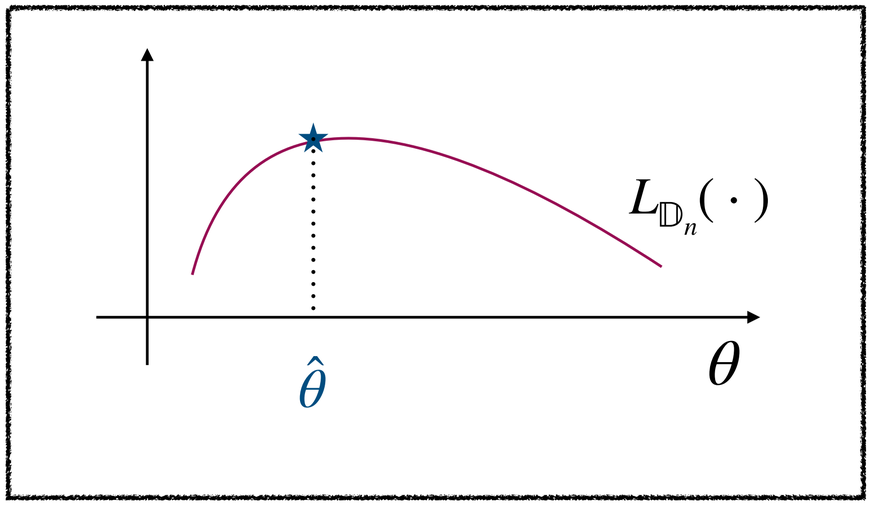
Likelihood function captures how likely it is to have observed $\mathbb{D}_n$ under parameter θ

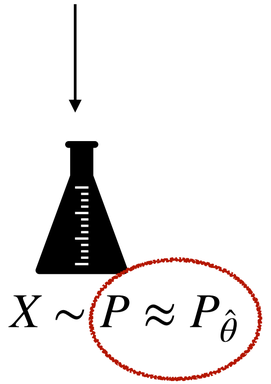
Statistical Problem: **Maximum Likelihood Estimator**

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L_{\mathbb{D}_n}(\theta)$$

Only if $Supp(L_{\mathbb{D}_n})$ independent of θ

$$\equiv \quad \hat{\theta} = \arg\max_{\theta \in \Theta} log(L_{\mathbb{D}_n}(\theta)) \quad \equiv \quad \hat{\theta} = \arg\min_{\theta \in \Theta} - log(L_{\mathbb{D}_n}(\theta))$$



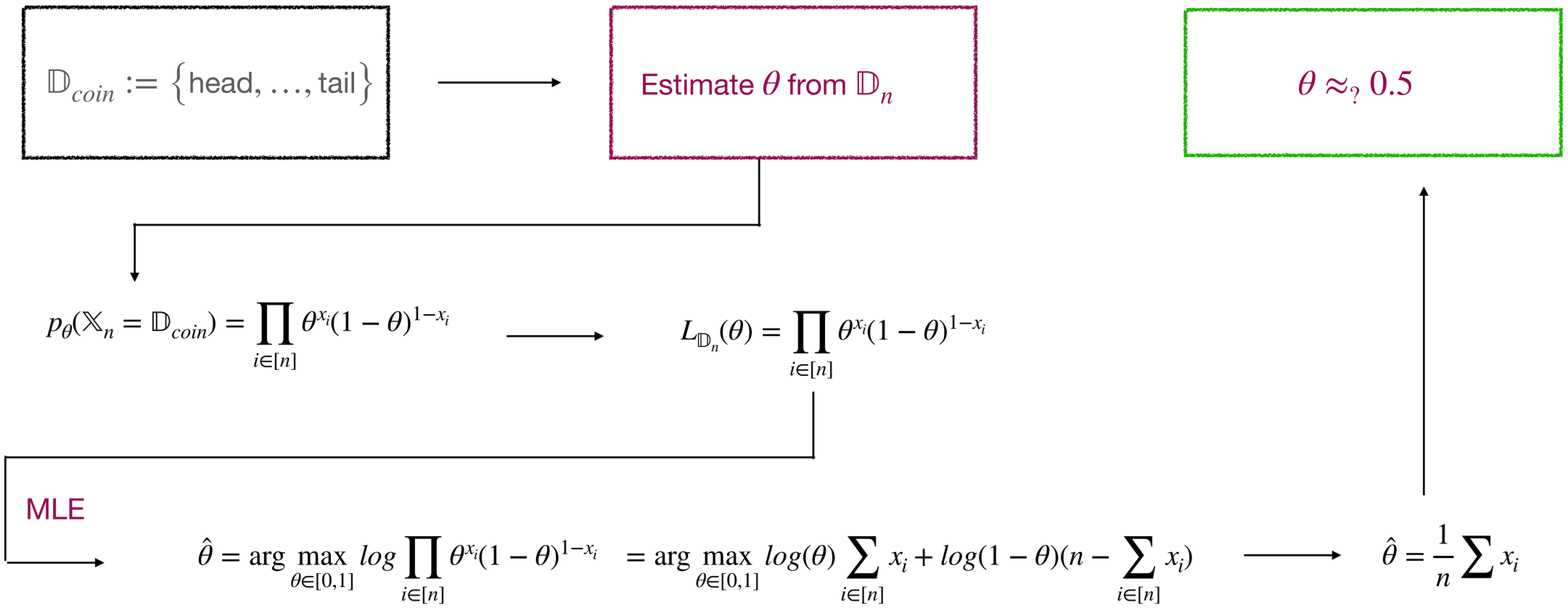


Statistics → we **estimated** the underlying process

≡

Machine Learning → we **learned** the underlying process

Statistical Problem: A Parametric Model : Tossing Example (continued)



Bayesian VS Frequentist: A Different Philosophy of Probabilistic Modelling

Frequentist Postulates VS Bayesian Postulates

All of Statistics: Chapter 11

https://egrcc.github.io/docs/math/all-of-statistics.pdf

Bayesian Method: Generic Mechanism

Probability over parameter space Θ

1 $\theta \sim p(\theta) =:$ **prior** distribution

Choice of statistical model $p(x \mid \theta)$

(belief about data x given parameter θ)

2 $X \sim p(x \mid \theta)$

Given data set of n iid realisations

update our belief

3 $p(\theta \mid \mathbb{D}_n) =:$ **posterior** distr.

Data Set :

$\mathbb{D}_n = \{X_i\}_{i \in [n]} \quad X_i \sim_{iid} p(X \mid \theta)$

$$p(\mathbb{D}_n \mid \theta) = \prod_{i \in [n]} p_\theta(x_i) = L_{\mathbb{D}_n}(\theta)$$

Likelihood distr.

Plug-in

Baye's Theorem

$$p(\theta \mid x) = \frac{p(x \mid \theta) p(\theta)}{\int_\Theta p(x \mid \theta) p(\theta)}$$

$$p(\theta \mid \mathbb{D}_n) = \frac{L_{\mathbb{D}_n}(\theta) p(\theta)}{\int_\Theta L_{\mathbb{D}_n}(\theta) p(\theta)}$$

ETH zürich

Bayesian Method: Point Estimate : MAP

$$p(\theta \mid \mathbb{D}_n) = \frac{L_{\mathbb{D}_n}(\theta)p(\theta)}{\int_\Theta L_{\mathbb{D}_n}(\theta)p(\theta)}$$

$$p(\theta \mid \mathbb{D}_n) \propto L_{\mathbb{D}_n}(\theta)p(\theta)$$

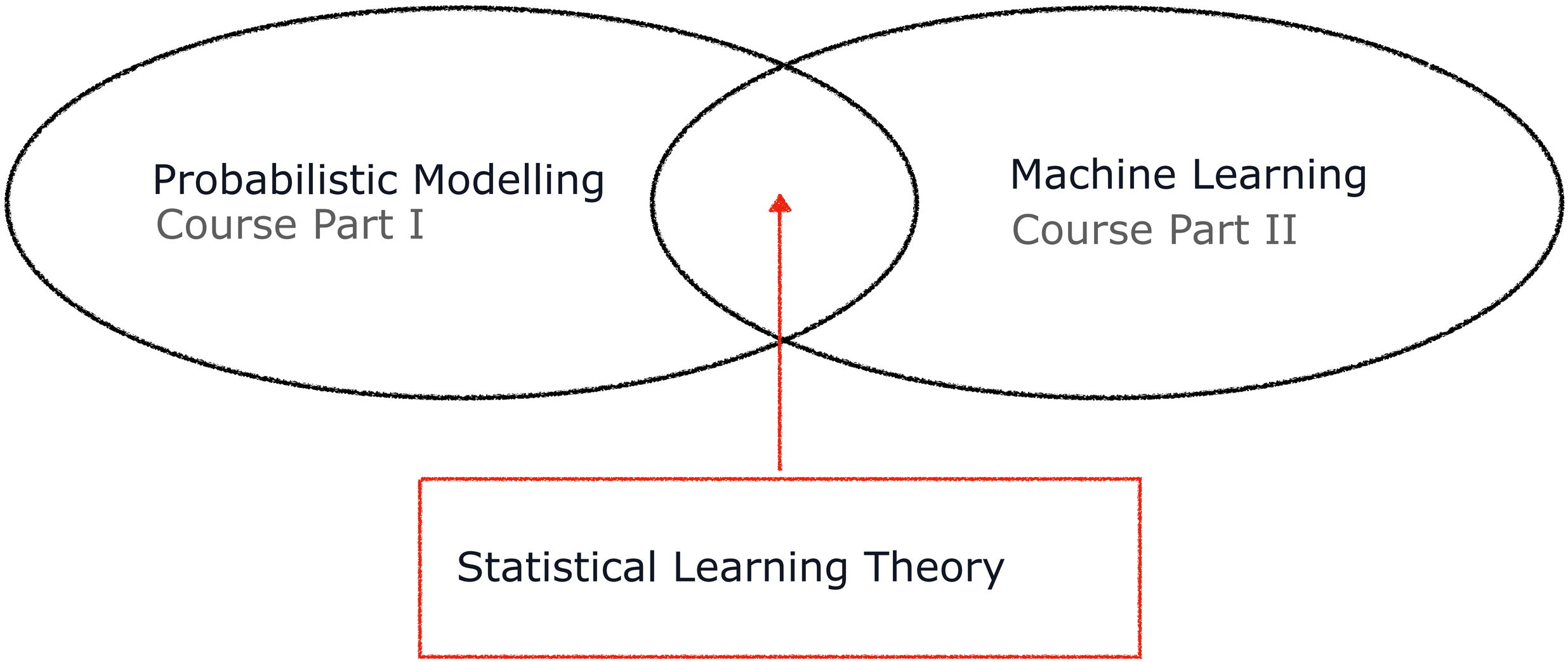
Posterior $\propto$ **Likelihood** $\times$ **Prior**

Maximum **A** **P**osteriori Estimate

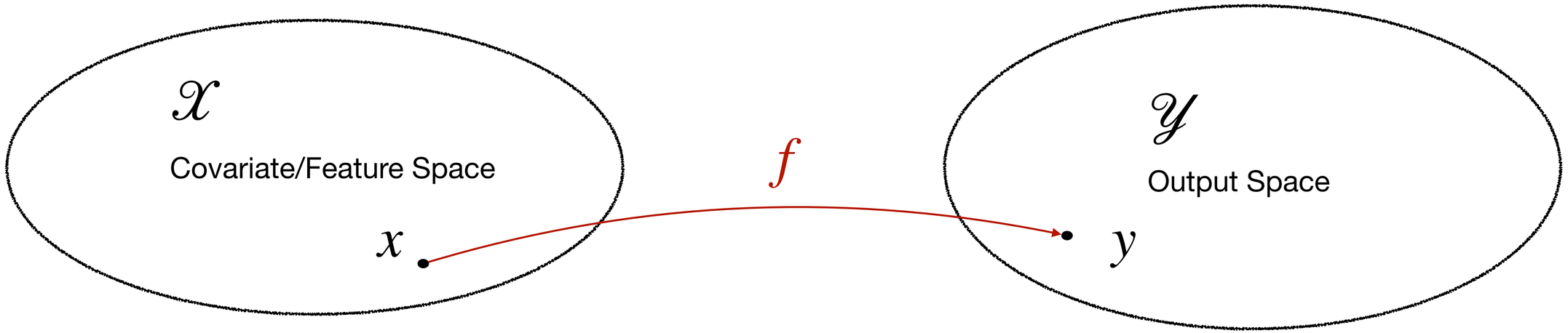
$$\hat{\theta}_{MAP} := \arg\max_{\theta \in \Theta} p(\theta \mid \mathbb{D}_n)$$

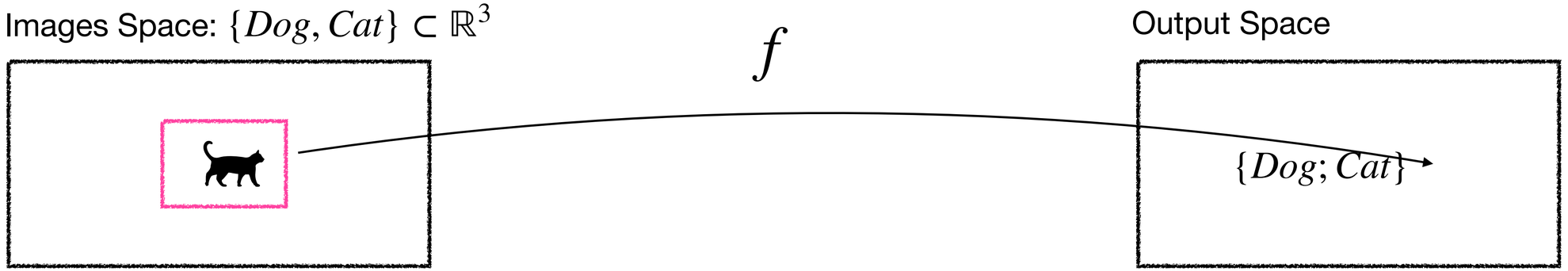
$$= \arg\max_{\theta \in \Theta} log(L_{\mathbb{D}_n}(\theta)) + log(p(\theta))$$ (under some regularity assumptions)

Overview

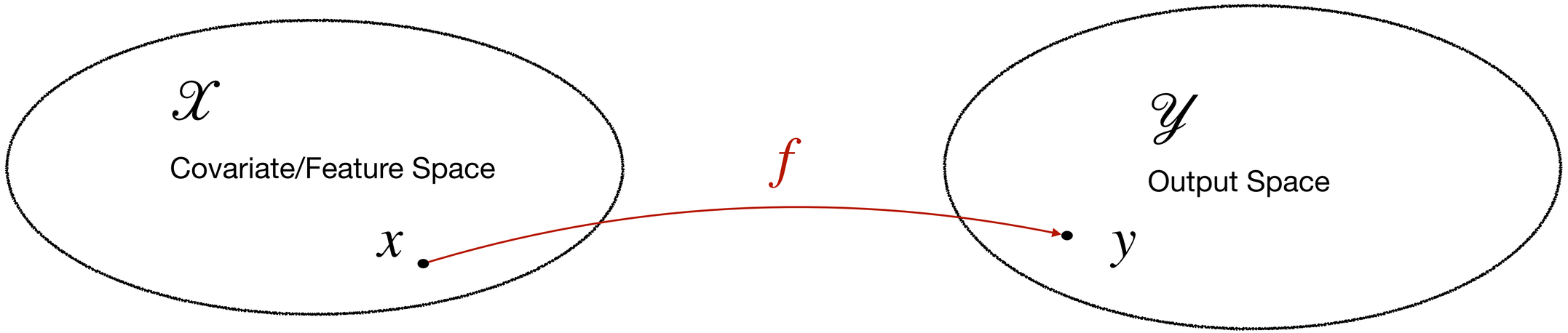

Probabilistic Modelling
Course Part I
Machine Learning
Course Part II
Statistical Learning Theory

From Stochastic to ML: **Fundamentals of Supervised Learning Model**





From Stochastic to ML: **Fundamentals of Supervised Learning Model**



" Supervised " $\Longrightarrow$ learn/ estimate/ approximate f from $\mathbb{D}_n = \{x_i, y_i\}_{i \in [n]}$

Additive Model:

$$y = f(x) + \epsilon$$

$\epsilon \sim P_\epsilon$ Random Perturbation

$f(x)$ Systematic Component

ETH zürich

From Stochastic to ML: Fundamentals of Supervised Learning Model

“ Supervised ” $\Longrightarrow$ learn/ estimate/ approximate from $\mathbb{D}_n = \{x_i, y_i\}_{i \in [n]}$

Additive Model (zero -mean Gaussian Assumption):

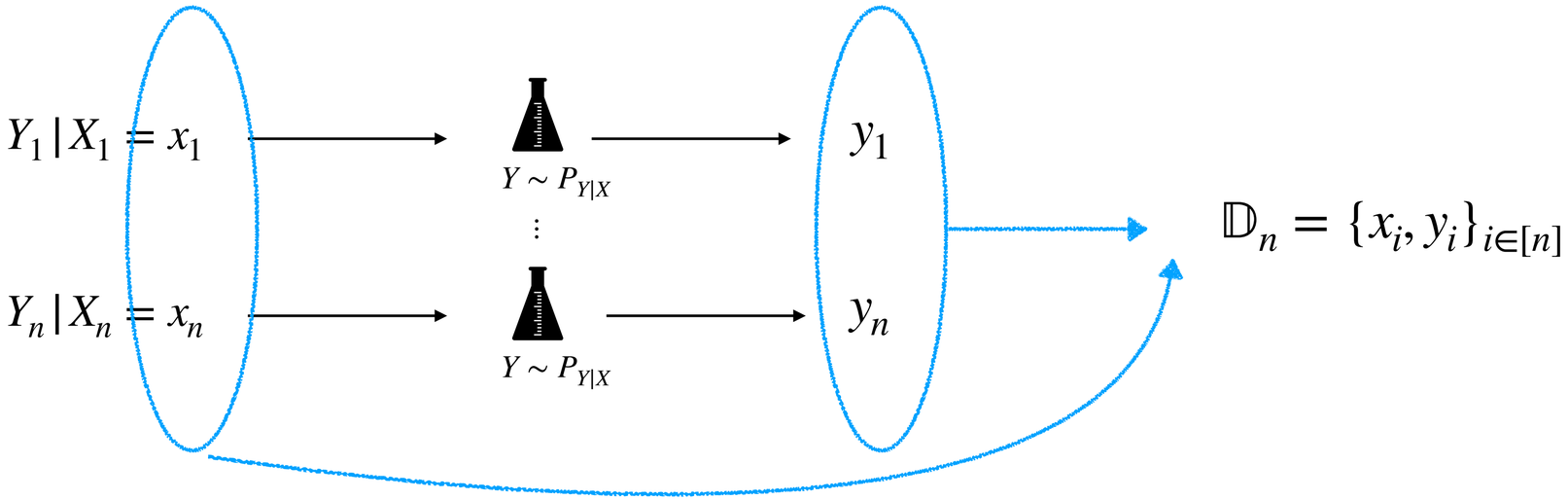
$$y = f(x) + \epsilon \qquad \epsilon \sim N(0,\sigma^2) \rightarrow y \sim N(f(x), \sigma^2)$$

Data Modelling (recall stochastic part !)

$$y \sim N(f(x), \sigma^2) \rightarrow$$ **knowing** $P_{Y|X} \Longrightarrow$ **knowing** $f(x)$

From Stochastic to ML: **Fundamentals of Supervised Learning Model**

Data Modelling (recall stochastic part !)



Note: the iid assumption is **key** to encode the fact that we assume similarity in the data generating process i.e the systematic component

From Stochastic to ML: Fundamentals of Supervised Learning Model

" Supervised " $\Longrightarrow$ learn/ estimate/ approximate from $\mathbb{D}_n = \{x_i, y_i\}_{i\in[n]}$

Additive Model (zero -mean Gaussian Assumption):

$$y = f(x) + \epsilon \qquad \epsilon \sim N(0,\sigma^2) \longrightarrow y \sim N(f(x), \sigma^2)$$

Model Function Class: Recall $P_{Y|X}$ unknown $\rightarrow \approx P_{Y|X;\theta}$

$$f \in \mathcal{F}_{model}$$

In this course: parametrised models

$$\mathcal{F}_{model} = \mathcal{F}_\theta = \{f(x;\theta) \,|\, \theta \in \Theta\}$$

Model Function Class: Examples

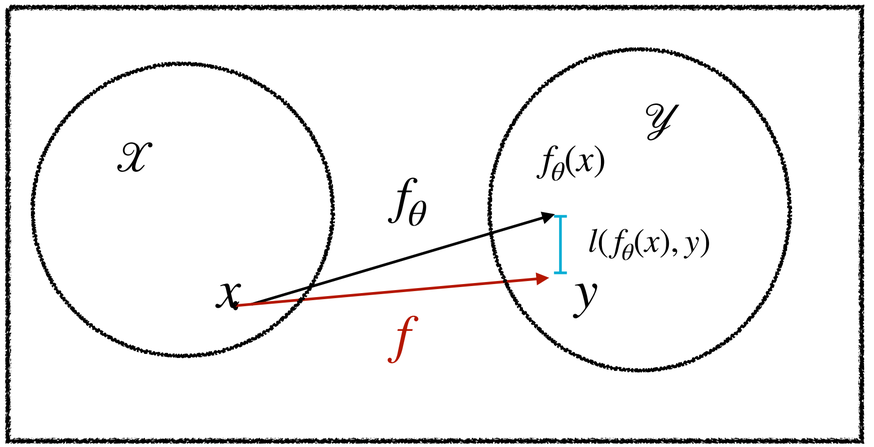
linear models $\quad \mathcal{F}_\theta = \{\langle x, \theta\rangle \,|\, \theta \in \Theta\}$

Neural Network $\quad \mathcal{F}_\theta = \{NN(x;\theta) \,|\, \theta \in \Theta = \{W_i, b_i\}_{i\in[L]}\} \longrightarrow W_i, b_i$ Weights of $NN :=$ Neural network

ETH zürich

From Stochastic to ML: Fundamentals of Supervised Learning Model

How to quantify the quality of a model function class given $\mathbb{D}_n = \{x_i, y_i\}_{i\in[n]}$ **?**



Loss function

$l : \Theta \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$

$l(f_\theta(X), Y)$

e.g square loss $\quad l(f_\theta(X), Y) = \|f_\theta(X) - Y\|^2$

Quality of loss **on average over the data space**

Loss function is a random object !

Risk function (population)

$$R(\theta) := \mathbb{E}_{X,Y} l(f_\theta(X), Y)$$

$P_{X,Y}$ Unknown

LLN

$\mathbb{D}_n = \{x_i, y_i\}_{i\in[n]}$

Empirical Risk function

$$\hat{R}(\theta) := \frac{1}{n} \sum_{i\in[n]} l(f_\theta(x_i), y_i)$$

Computable given the DATA !!!

From Stochastic to ML: Fundamentals of Supervised Learning Model

How to quantify the quality of a model function class given $\mathbb{D}_n = \{x_i, y_i\}_{i\in[n]}$ **?**

Risk function (population)

$$R(\theta) := \mathbb{E}_{X,Y} l(f_\theta(X), Y)$$

→ $P_{X,Y}$ Unknown

LLN $\mathbb{D}_n = \{x_i, y_i\}_{i\in[n]}$ →

Empirical Risk function

$$\hat{R}(\theta) := \frac{1}{n}\sum_{i\in[n]} l(f_\theta(x_i), y_i)$$

Computable given the DATA !!!

Empirical **R**isk **M**inimisation

$$\hat{\theta} := \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i\in[n]} l(f_\theta(x_i), y_i)$$

Prediction: given a new data point X_{new} **predict** Y_{new}

$$\hat{y}_{new} = f_{\hat{\theta}}(x_{new})$$

From Stochastic to ML: Fundamentals of Supervised Learning Model

How to quantify the quality of a model function class given $\mathbb{D}_n = \{x_i, y_i\}_{i\in[n]}$ **?**

Risk function (population)

$$R(\theta) := \mathbb{E}_{X,Y} l(f_\theta(X), Y)$$

$\longrightarrow$ $P_{X,Y}$ Unknown

LLN

$\mathbb{D}_n = \{x_i, y_i\}_{i\in[n]}$ $\longrightarrow$

Empirical Risk function

$$\hat{R}(\theta) := \frac{1}{n}\sum_{i\in[n]} l(f_\theta(x_i), y_i)$$

Computable given the DATA !!!

Empirical **R**isk **M**inimisation

$$\hat{\theta} := \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i\in[n]} l(f_\theta(x_i), y_i)$$

$n \to \infty$

$R(\theta) \approx \hat{R}(\theta)$

For some class of function $\mathscr{F}_\theta$

Notice: the more data we have, the closer we should be from the true quantitative analysis from the model we defined

Statistical Problem: Example: Linear Regression (fundamental ml modelling)

Data Modelling Additive Assumption:

$$y = f(x) + \epsilon \quad \epsilon \sim N(0,\sigma^2) \qquad x \in \mathbb{R}^d \quad , \quad y \in \mathbb{R}$$

Model Function Class: Recall $P_{Y|X}$ unknown $\rightarrow \approx P_{Y|X;\theta}$

$$\mathcal{F}_\theta = \{x^T\theta \quad | \quad \theta \in \mathbb{R}^d\}$$

Loss function : quadratic loss

$$l(\langle\theta, x\rangle, y) = \|x^T\theta - y\|^2$$

Find best θ given data set $\mathbb{D}_n$

$$\hat{\theta} = \arg\min \frac{1}{n} \sum \|x_i^T\theta - y_i\|^2 = \|\tilde{X}\theta - \tilde{y}\|^2 \qquad \text{With} \qquad \tilde{X} = [x_1^T, \ldots, x_n^T]^T \quad \tilde{y} = [y_1, \ldots, y_n]^T$$

Data Modelling Parametric Assumption

$$y = \langle x, \theta\rangle + \epsilon \quad \epsilon \sim N(0,\sigma^2) \implies y \,|\, x \sim N(x^T\theta, \sigma^2)$$

Why "Regression " ?

$\mathbb{E}\{Y|X = x\} =:$ Regression function

$$\mathbb{E}\{Y|X = x\} = x^T\theta \xrightarrow{\hat{\theta}} \mathbb{E}\{Y|X = x\} \approx x^T\hat{\theta}$$

We learned/approximate the regression function !

Statistical Problem: Linear Regression: Least Square Estimation

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \|\tilde{X}\theta - \tilde{y}\|^2 \qquad \text{With} \qquad \tilde{X} = [x_1^T, \ldots, x_n^T]^T \quad \tilde{y} = [y_1, \ldots, y_n]^T$$

LS solution (linear algebra)

$$\hat{\theta} = (\tilde{X}^T\tilde{X})^+\tilde{X}^T\tilde{y}$$

Prediction

$$\hat{y}_{new} = x_{new}^T\hat{\theta}$$

Recall: additive model

$$y = f_\theta(x) + \epsilon$$

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \|\epsilon\|^2$$

Minimum error in l2 norm with respect to our model assumption !

Statistical Problem: Linear Regression: MLE

Additive Model (zero -mean Gaussian Assumption):

$$y = \langle x, \theta \rangle + \epsilon \quad \epsilon \sim N(0,\sigma^2) \implies y \,|\, x \sim N(x^T\theta, \sigma^2)$$

Unknown parameter :

$$y \,|\, x \sim N(x^T\theta, \sigma^2)$$

Estimating the Mean and Variance MLE

$$\{\hat{\theta}_{MLE}, \sigma_{MLE}\} = \arg \min_{\theta \in \mathbb{R}^d, \sigma \in \mathbb{R}} - log(L_{\mathbb{D}_n}(\theta; \sigma))$$

$$\longrightarrow \quad L_{\mathbb{D}_n}(\theta) = (2\pi\sigma^2)^{-n/2} exp(-\frac{1}{2\sigma^2} \sum_{i \in [n]} (y_i - x_i^T\theta)^2)$$

$$\longrightarrow \quad log(L_{\mathbb{D}_n}(\theta)) \propto -\frac{n}{2} ln(\sigma^2) - \frac{1}{2\sigma^2} \|\tilde{y} - \tilde{X}\theta\|^2$$

$$\hat{\theta}_{MLE} = \hat{\theta}_{LS} = (\tilde{X}^T\tilde{X})^+ \tilde{X}^T\tilde{y}$$

$$\sigma^2_{MLE} = \frac{1}{n} \sum_{i \in [n]} (y_i - x_i^T\hat{\theta}_{MLE})^2$$

$$\hat{y}_{new} = x_{new}^T \hat{\theta}_{MLE}$$

Confidence Interval

Prediction Interval

Statistical Problem: **Bayesian Linear Regression: MAP**

Additive Model (zero -mean Gaussian Assumption):

$$y = \langle x, \theta \rangle + \epsilon \quad \epsilon \sim N(0,\sigma^2), \quad \theta \sim p(\theta) \implies y \,|\, x, \theta \sim N(x^T\theta, \sigma^2)$$

Recall: Prior user defined !

$$p(\theta)$$

Choose a Gaussian Prior

$$p(\theta) = N(0,\sigma_p)$$

MAP Estimate

$$\theta_{MAP} = \arg\max_{\theta} p(\theta \,|\, \mathbb{D}_n) = log(L_{\mathbb{D}_n}) + log(p(\theta))$$
$$= \arg\min_{\theta} \frac{1}{\sigma_n^2} \sum_{i \in [n]} (y_i - \theta^T x_i)^2 + \frac{1}{\sigma_p^2} \|\theta\|^2$$
$$= \arg\min_{\theta} \|Y - X\theta\|^2 + \frac{\sigma_n^2}{\sigma_p^2} \|\theta\|^2$$

$$\theta_{MAP} = (X^T X + \frac{\sigma_n^2}{\sigma_p^2} \mathbb{I})^{-1} X^T y$$

$\equiv$ Ridge Regression (or L_2 constraint least squares)

Statistical Problem: **Bayesian Linear Regression: MAP Prior Choice**

Prior

$p(\theta)$

Gaussian

$$logp(\theta) \propto -\frac{1}{2\sigma_p^2}\|\theta\|_2^2$$

L_2 **Regularisation (Ridge)**

$$\theta_{MAP} = \arg\min_{\theta} \|Y - X\theta\|^2 + \frac{\sigma_n^2}{\sigma_p^2}\|\theta\|^2$$

Laplacian

$$logp(\theta) \propto -\frac{1}{\sigma_p}\|\theta\|_1^2$$

L_1 **Regularisation (LASSO)**

$$\theta_{MAP} = \arg\min_{\theta} \|Y - X\theta\|^2 + \frac{1}{\sigma_p}\|\theta\|_1^2$$

Statistical Problem: **Bayesian Modelling : Conjugate Prior**

Definition :

When the **prior** and the **posterior** are from the same family of distribution we say that the prior is **conjugate** with respect to the model

Special Note :

Data Gaussian + Prior Gaussian $\Longrightarrow$ Gaussian Posterior

Closed Form Posterior

Analytical Tractability

Useful for :

Incremental Learning

Predictive distribution

Statistical Problem: Bayesian Modelling : Conjugate Prior for Gaussian Data

Recall Linear Gaussian Noise Additive Model :

$$L_{\mathbb{D}_n}(\theta) = \prod_{i \in [n]} p(y_i \,|\, x_i, \theta) \quad p(y_i \,|\, x_i \theta) = N(\theta^T x_i, \sigma_n)$$

Zero-Mean Gaussian Prior:

$$p(\theta) = N(0, \sigma_p)$$

Is Posterior Gaussian ? i.e $p(\theta)$ conjugate prior ?

$$p(\theta \,|\, \mathbb{D}_n) = L_{\mathbb{D}_n}(\theta) p(\theta) = \prod N_y(\theta^T x_i, \sigma_n) N_\theta(0, \sigma_p) \quad \Leftrightarrow \quad \text{Qty Gaussian in } \theta$$

Next slide: Simplified proof for univariate gaussian !

Product of Univariate Gaussian

Note: For Multivariate, Results on Gaussian Conditioned on Gaussian are used to for the proof

$$p(\theta \mid \mathbb{D}_n) = L_{\mathbb{D}_n}(\theta)p(\theta) = \prod N_y(\theta^T x_i, \sigma_n)N_\theta(0,\sigma_p) = \frac{1}{(2\pi\sigma_n)^{n/2}(2\pi\sigma_p)^{1/2}} exp\left\{ (-1)(\underbrace{\frac{1}{2\sigma_n^2}\sum_{i\in[n]}(y_i - x_i\theta)^2}_{:=\beta} + \frac{1}{2\sigma_p^2}\theta^2) \right\}$$

$$\beta = \frac{1}{2\sigma_n^2}\sum_{i\in[n]}(y_i - x_i\theta)^2 + \frac{1}{2\sigma_p^2}\theta^2 = \frac{1}{2\sigma_n^2}(\sum_i y_i^2 - 2\theta\sum_i x_i y_i + \theta^2\sum_i x_i^2) + \frac{1}{2\sigma_p^2}\theta^2$$

$$\sum_i y_i^2 := y_n^2$$

$$\sum_i y_i x_i := y x_n$$

$$\sum_i x_i^2 := x_n^2$$

$$\beta = \frac{\theta^2(\sigma_n^2 + \sigma_p^2 x_n^2) - 2\theta y x_n \sigma_p^2 + \sigma_p^2 y_n^2}{2\sigma_p^2\sigma_n^2} = \frac{\theta^2 - 2\theta(\frac{y x_n \sigma_p^2}{\sigma_n^2 + \sigma_p^2 x_n^2}) + \frac{\sigma_p^2 y_n^2}{\sigma_n^2 + \sigma_p^2 x_n^2}}{\frac{2\sigma_p^2\sigma_n^2}{\sigma_n^2 + \sigma_p^2 x_n^2}}$$

$$=: \frac{\theta^2 - 2A\theta + B}{C}$$

Product of Univariates Gaussian

Define:

$$\epsilon = \frac{A^2 - A^2}{C} = 0$$

Then :

$$\beta = \beta + \epsilon = \frac{(\theta - A)^2}{C} + \frac{A^2 - B}{C}$$

NOTE: What did we show ?

Proportionality !

For valid density: need to workout the scaling factor !

Which implies :

Scaling Factor

$$exp\left\{ -\beta \right\} = exp\left\{ -\frac{(\theta - A)^2}{C} \right\} exp\left\{ \frac{A^2 - B}{C} \right\}$$

Recall what we aim :

$$p(\theta \mid \mathbb{D}_n) \propto exp\left\{ -\frac{(\theta - \mu_\theta)^2}{2\sigma_\theta^2} \right\}$$

$$\mu_\theta = A = \frac{yx_n\sigma_p^2}{\sigma_n^2 + \sigma_p^2 x_n^2}$$

$$\sigma_\theta^2 = \frac{1}{2}C = \frac{\sigma_p^2\sigma_n^2}{\sigma_n^2 + \sigma_p^2 x_n^2}$$

Bayes Classifier : **Framework**

Define:

Features : $X \in \mathbb{R}^d$

Output : $Y \in \{1, \ldots, C\} \subset \mathbb{N}$

Classifier :

$$h : X \to Y$$

Definition: Baye's Classifier

$$C_{Bayes}(x) = \arg \max_{c \ in[C]} P(Y = c \,|\, X = x)$$

Baye's Classifier Optimal wrt Setting !

Intuition: (not a Proof !) Recall Risk Framework

Error of Misclassification :

$$P(Y \neq h(X)) = \mathbb{E}[\![_{Y \neq h(X)} = \mathbb{E}_X \mathbb{E}_{Y|X}[\![_{Y \neq h(X)|X=x}$$

$$= \mathbb{E}_X P(Y = 1 \,|\, x)[\![_{h(x)=2,\ldots,C} + P(Y = 2 \,|\, x)[\![_{h(x)=1,3,\ldots,C} \cdots$$

Minimised by maximising the chance of not making an error !

Bayes Classifier : **Naive Baye's Classifier**

Definition: Baye's Classifier

$$C_{Bayes}(x) = \arg\max_{c \ in[C]} P(Y = c \,|\, X = x)$$

Assume $X = \{x_i\}_{i \in [n]}$ iid then the Baye's Classifier yield the **naive** baye's classifier, where naive stand for the assumption of the independence in the features

$$C^{Naive}_{Bayes}(x) = \arg\max_{c \ in[C]} P(Y = c) \prod_{i \in [n]} P(x_i)$$