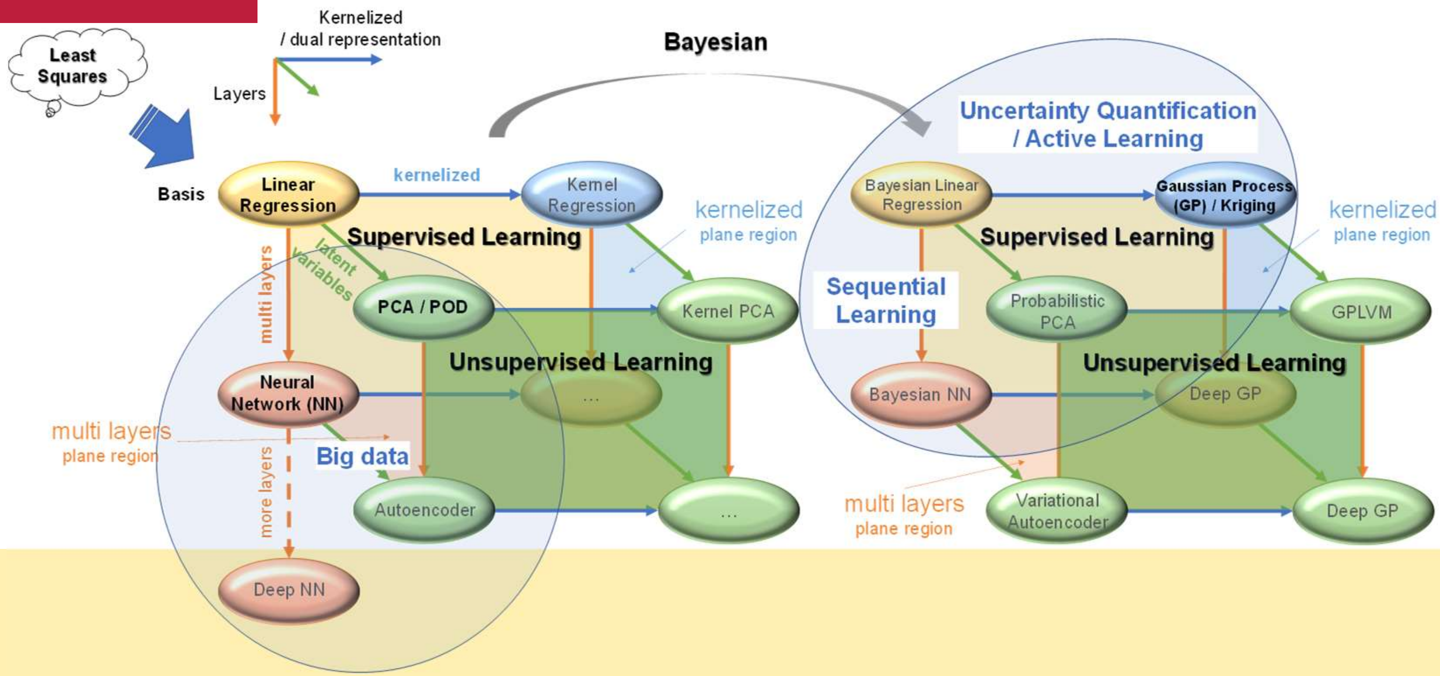



# Scientific Machine Learning

*Lecture 2: Curve Fitting and Probability Theory*

Dr. Daigo Maruyama

Prof. Dr. Ali Elham

# Lecture content

- Polynomial Curve Fitting

- Probability Theory

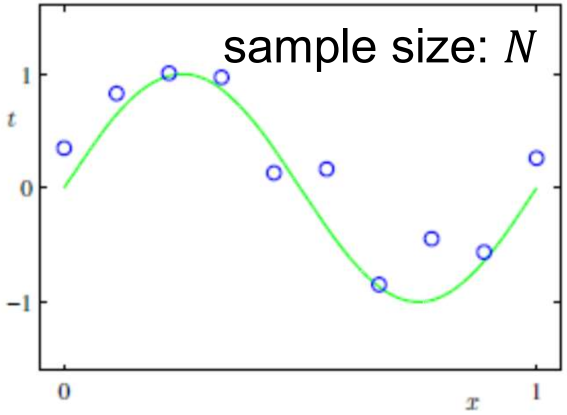
The lecture of this time basically follows the 1st chapter of the book:
Christopher M. Bishop "Pattern Recognition And Machine Learning" Springer-Verlag (2006)
The name of this book is shown as “PRML” when it is referred in the slides.

## Lecture content

1. Polynomial Curve Fitting

# Polynomial Curve Fitting

Example: The following data points are given.
Which kind of functions do you want to put on them?




PRML, p. 4

Assumption:
The data points lie on a cubic function.

$$y(x, \boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 = \sum_{i=0}^{3} w_i x^i$$

**Linear model**: a linear function of <u>the coefficients $\boldsymbol{w}$</u>

**parameters**

**Error function**

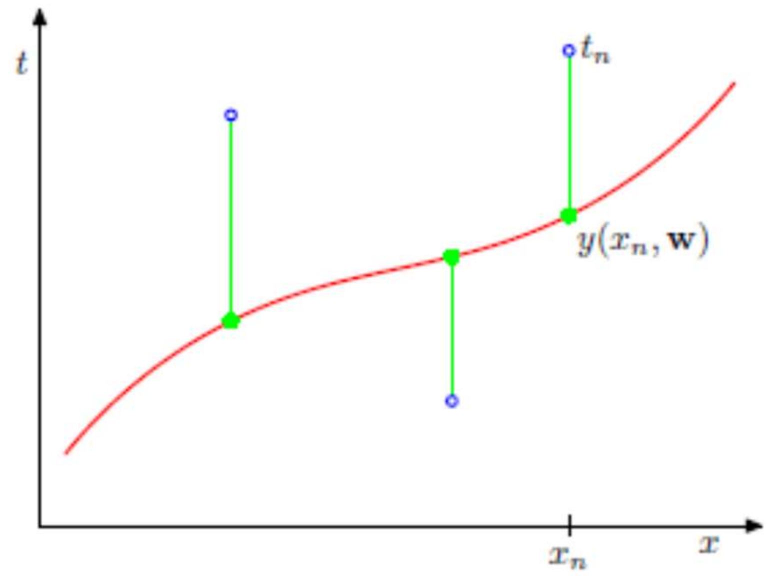
$$E(\boldsymbol{w}) = \sum_{i=1}^{N} \{t_i - y(x_i, \boldsymbol{w})\}^2$$

**Least squares method itself**

$$\boldsymbol{w}^* = \underset{\boldsymbol{w}}{\operatorname{argmin}}\, E(\boldsymbol{w})$$

# Polynomial Curve Fitting

## Visualizing the least square method




PRML, p. 6

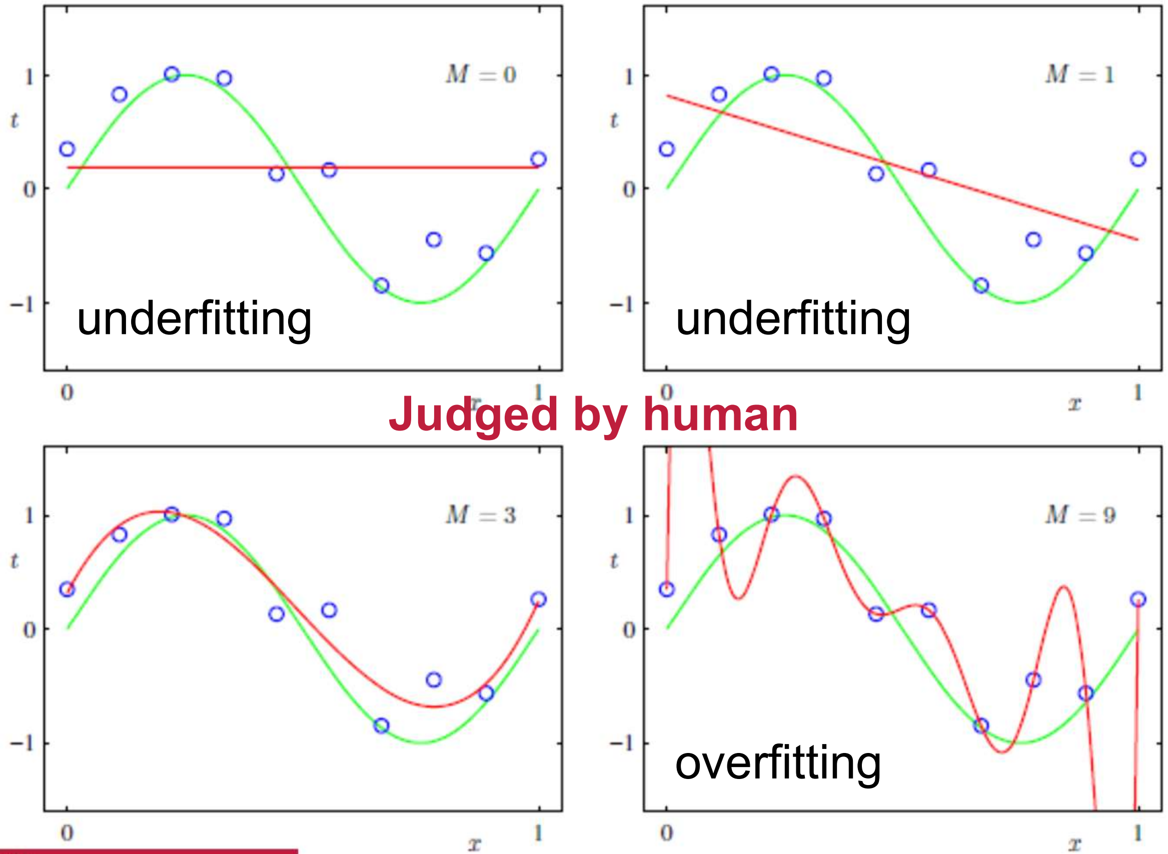
$$E(\boldsymbol{w}) = \sum_{i=1}^{N} \{t_i - y(x_i, \boldsymbol{w})\}^2$$

# Polynomial Curve Fitting

various orders $M$

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{i=0}^{M} w_i x^i$$

sample size: $N = 10$ (common for all)




PRML, p. 7

Higher order models are more flexible but not always better for prediction.

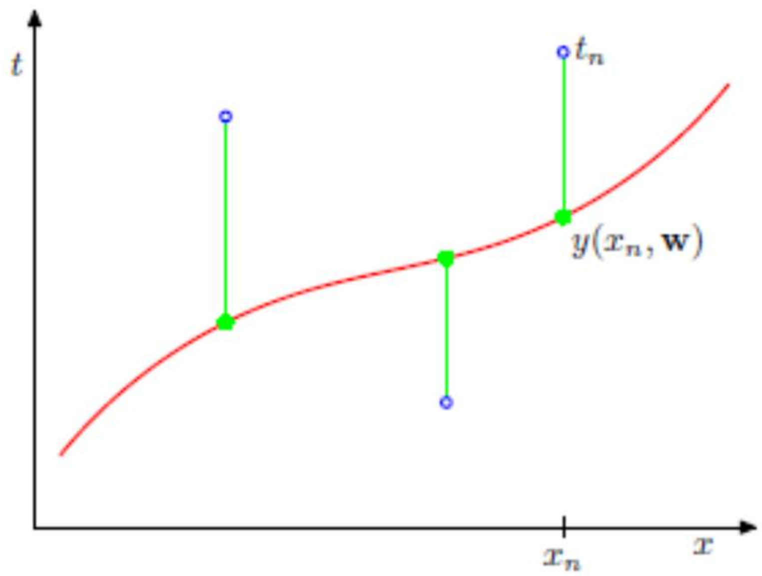
Selection of $M$ :
**Model selection**

Good prediction
= Good generalization

How to evaluate good **generalization** quantitatively?

# Polynomial Curve Fitting

## Least squares




PRML, p. 6

Quantitative Insight:

**root-mean-square error (RMS error)**

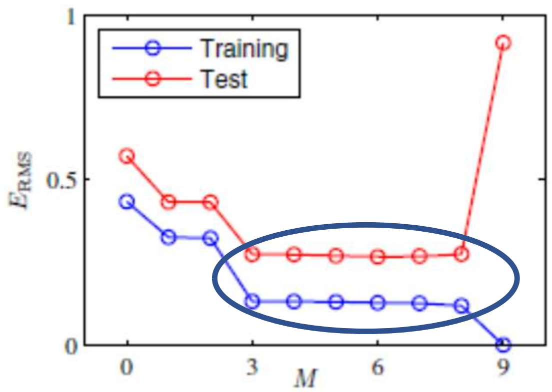
$$E_{RMS} = \sqrt{\frac{E(\boldsymbol{w})}{N}}$$

$$= \sqrt{\frac{1}{N}\sum_{i=1}^{N}\{t_i - y(x, \boldsymbol{w})\}^2}$$

Generalized by

- deleting the effect of sample size $N$
- using the same unit as the output

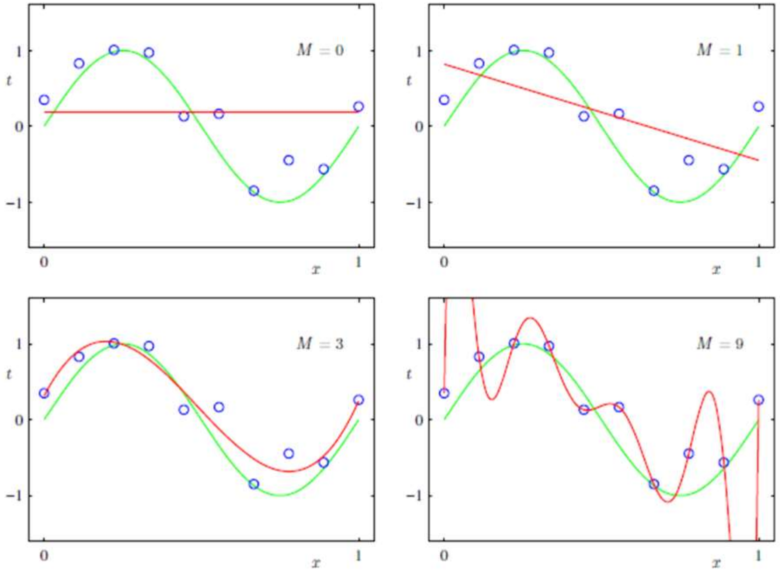
# Polynomial Curve Fitting




| | $M=0$ | $M=1$ | $M=6$ | $M=9$ |
|---|---|---|---|---|
| $w_0^\star$ | 0.19 | 0.82 | 0.31 | 0.35 |
| $w_1^\star$ | | -1.27 | 7.99 | 232.37 |
| $w_2^\star$ | | | -25.43 | -5321.83 |
| $w_3^\star$ | | | 17.37 | 48568.31 |
| $w_4^\star$ | | | | -231639.30 |
| $w_5^\star$ | | | | 640042.26 |
| $w_6^\star$ | | | | -1061800.52 |
| $w_7^\star$ | | | | 1042400.18 |
| $w_8^\star$ | | | | -557682.99 |
| $w_9^\star$ | | | | 125201.43 |

PRML, p. 6

${E_{RMS}}^{(Training)} = 0$ at $M = 9$
but ${E_{RMS}}^{(Test)}$ is large.




sample size
$N = 10$
(common for all)

**Degree of freedom** of the polynomial function

$$y(x, \boldsymbol{w}) = \sum_{i=0}^{M} w_i x^i$$

Is equal to $N = 10$
when $M = 9$ (0,…,9)

# Polynomial Curve Fitting (Regularization)

Concept: constraints on the parameters $\boldsymbol{w}$ Regularization

$$\min_{\boldsymbol{w}} E(\boldsymbol{w})$$

s.t. $\|\boldsymbol{w}\|^2 \leq \eta$

The concept is equivalent to minimizing the following modified error function:

$$\min_{\boldsymbol{w}} E(\boldsymbol{w}) \qquad \text{where,} \quad E_{reg}(\boldsymbol{w}) = E(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|^2$$

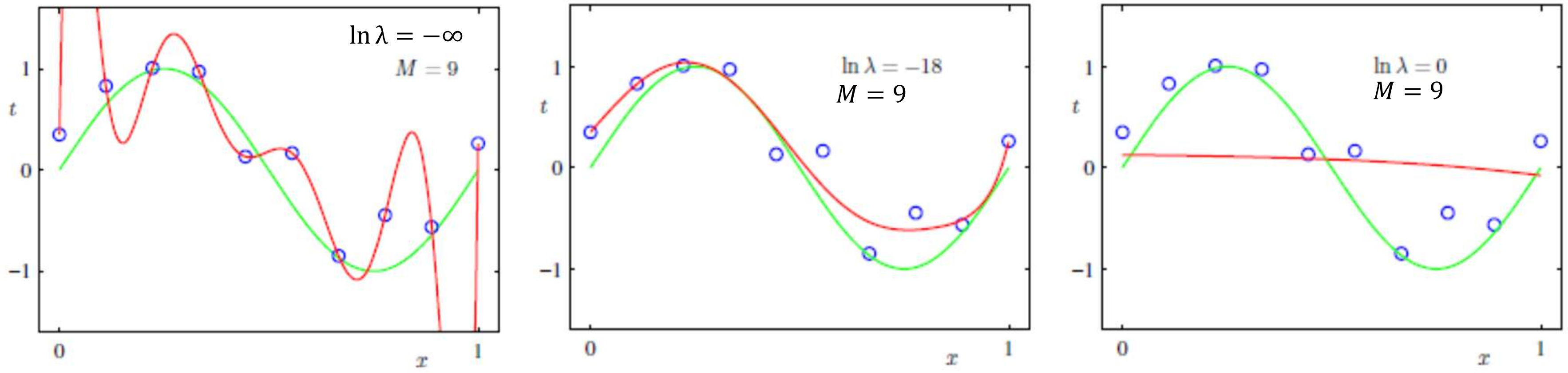
regularization term

$\lambda$ : A parameter to make a balance between the two terms:

- least square term
- regularization term

$$\|\boldsymbol{w}\|^2 = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} = w_0{}^2 + w_1{}^2 + \cdots + w_M{}^2$$

# Polynomial Curve Fitting (Regularization)

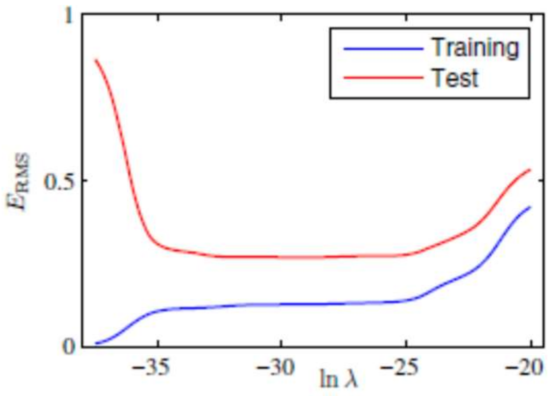



PRML, p. 10

We already know a technique of how to select $\lambda$.

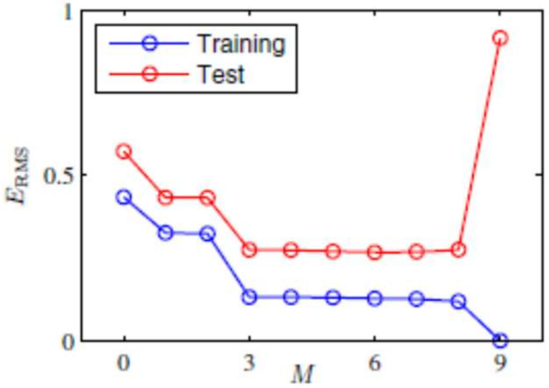
| | $\ln\lambda = -\infty$ | $\ln\lambda = -18$ | $\ln\lambda = 0$ |
|---|---|---|---|
| $w_0^\star$ | 0.35 | 0.35 | 0.13 |
| $w_1^\star$ | 232.37 | 4.74 | -0.05 |
| $w_2^\star$ | -5321.83 | -0.77 | -0.06 |
| $w_3^\star$ | 48568.31 | -31.97 | -0.05 |
| $w_4^\star$ | -231639.30 | -3.89 | -0.03 |
| $w_5^\star$ | 640042.26 | 55.28 | -0.02 |
| $w_6^\star$ | -1061800.52 | 41.32 | -0.01 |
| $w_7^\star$ | 1042400.18 | -45.95 | -0.00 |
| $w_8^\star$ | -557682.99 | -91.53 | 0.00 |
| $w_9^\star$ | 125201.43 | 72.68 | 0.01 |

PRML, p. 11




PRML, p. 11

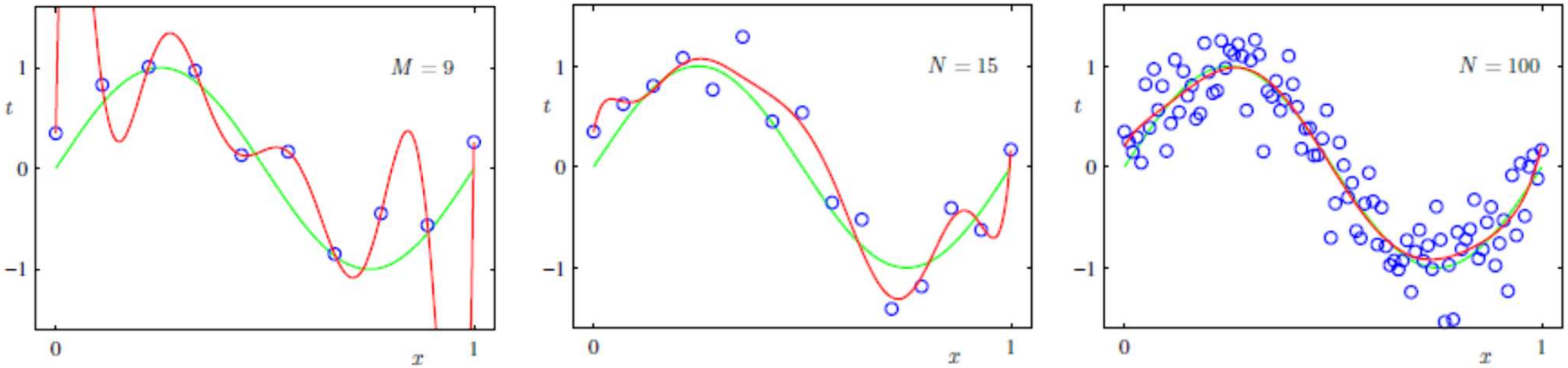
# Polynomial Curve Fitting




$$E_{RMS} = \sqrt{\frac{E(\boldsymbol{w})}{N}}$$

Training: $N = 10$
Test: $N$ is arbitrary

$M = 9$ for all (the same function)




PRML, p. 9

# Lecture content

2. Probability Theory

# Probability Theory

$x$: deterministic variable
$y$: random variable (or stochastic variable)

$$y \sim p(y|x)$$

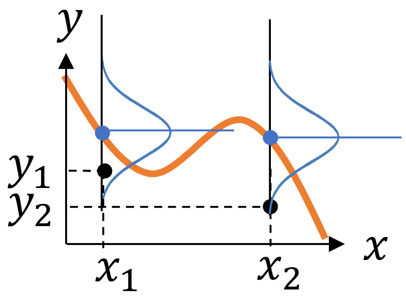
We pick up one of the observable quantities.

$$x \longrightarrow \boxed{p(y|x)} \longrightarrow y$$

$x$: input $\qquad$ $y$: output

**probability of $y$ when $x$ given**

**conditional probability**




$x$ can be also a probability.

## Probability Theory

A case when both $X$ and $Y$ are (discrete) random variables

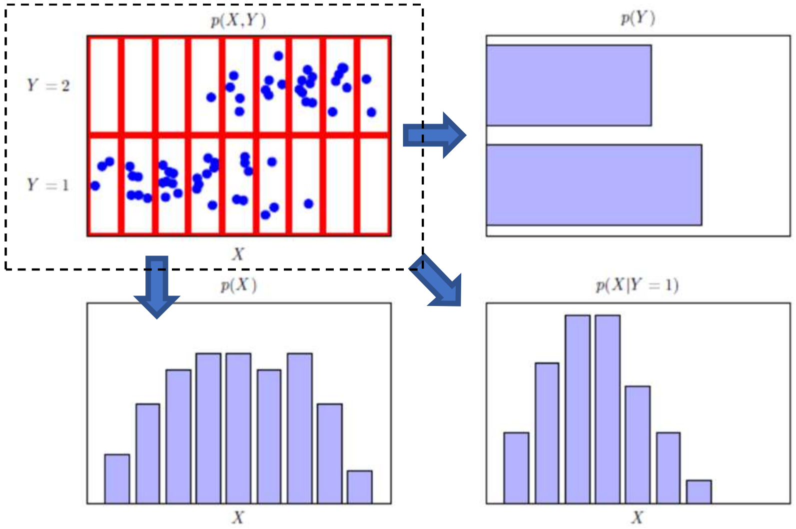
PRML, p. 16

$X$: $\{x_i\}, (i = 1, \cdots, 9)$
$Y$: $\{y_i\}, (i = 1,2)$

Classification in which category
sample size $N = 60$

histogram

**joint probability**
$P(X, Y)$




**marginal probability**
$P(Y)$
pretend that we did not see $X$

**marginal probability**
$P(X)$
pretend that we did not see $Y$

conditional probability
$P(X|Y = 1)$

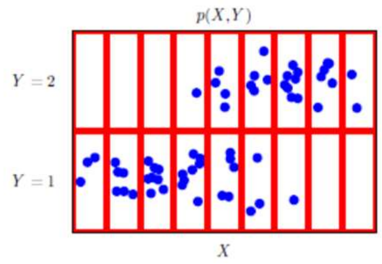
PRML, p. 16

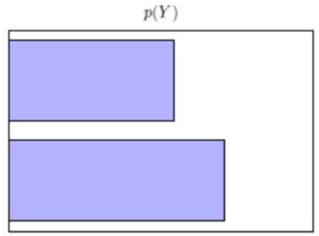
Joint probability contains all the information!

one of the goals in machine learning processes

# Probability Theory (Rules of Probability)

**joint probability**
$P(X,Y)$



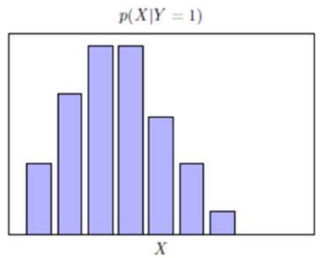




**marginal probability**

$$P(Y) = \sum_X P(X,Y)$$

**sum rule**

marginal distribution itself




$$\sum_X \sum_Y P(X,Y) = 1$$

$$\sum_X P(X, Y=1) = 1$$

$$P(X, Y=1) = P(X|Y=1)P(Y=1)$$

$$P(X,Y) = P(X|Y)P(Y)$$

**product rule**

# Probability Theory (Rules of Probability)

**sum rule** $P(Y) = \sum_X P(X, Y)$

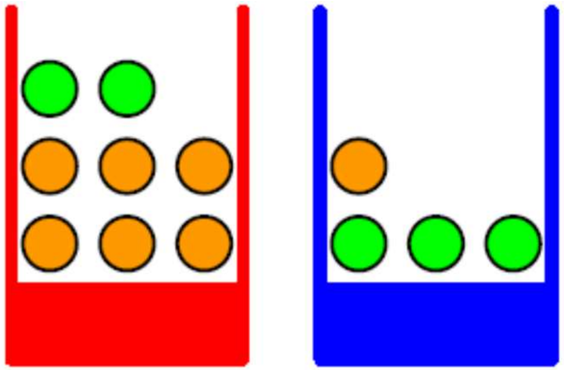
**product rule** $\underbrace{P(X, Y) = P(X|Y)P(Y) \qquad P(X, Y) = P(Y|X)P(X)}$

**Bayes' theorem** $P(Y|X) = \dfrac{P(X|Y)P(Y)}{P(X)}$

Interpretation is important.

Let us consider
**time flow / causality**

# Probability Theory (Introduction of Bayes' Theorem)



$$P(X = \text{"red box"}) = 0.6$$
$$P(X = \text{"blue box"}) = 0.4$$

1. First event $X$: choose one box
2. Second event $Y$: choose one piece of fruits

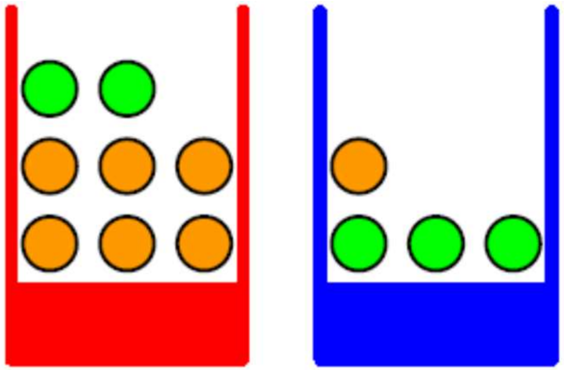
$$P(Y = \text{"orange"} | X = \text{"red box"}) = \frac{6}{8}$$

Then, $P(X = \text{"red box"} | Y = \text{"orange"}) = ?$

| | $P(X = \text{"red box"})$ | $P(X = \text{"blue box"})$ | $1$ |
|---|---|---|---|
| $P(Y = \text{"orange"})$ | $0.6 \times \frac{6}{8} = \frac{9}{20}$ | $0.4 \times \frac{1}{4} = \frac{1}{10}$ | $\frac{11}{20}$ |
| $P(Y = \text{"apple"})$ | $0.6 \times \frac{2}{8} = \frac{3}{20}$ | $0.4 \times \frac{3}{4} = \frac{3}{10}$ | $\frac{9}{20}$ |
| $1$ | $0.6$ | $0.4$ | $1$ |

Obtain joint probability: $P(X, Y)$

# Probability Theory (Introduction of Bayes' Theorem)



$P(X = \text{"}red\ box\text{"}) = 0.6$
$P(X = \text{"}blue\ box\text{"}) = 0.4$

## Bayes' theorem

$$P(X = \text{"}r\text{"} | Y = \text{"}o\text{"}) = \frac{P(Y = \text{"}o\text{"} | X = \text{"}r\text{"}) P(X = \text{"}r\text{"})}{P(Y = \text{"}o\text{"})}$$

$$= \frac{P(X = \text{r}, Y = \text{"}o\text{"})}{P(Y = \text{"}o\text{"})} = \frac{\frac{9}{20}}{\frac{11}{20}} = \frac{9}{11}$$

somehow understandable

**time flow / causality: reverse**

| | $P(X = \text{"}red\ box\text{"})$ | $P(X = \text{"}blue\ box\text{"})$ | $1$ |
|---|---|---|---|
| $P(Y = \text{"}orange\text{"})$ | $0.6 \times \frac{6}{8} = \frac{9}{20}$ | $0.4 \times \frac{1}{4} = \frac{1}{10}$ | $\frac{11}{20}$ |
| $P(Y = \text{"}apple\text{"})$ | $0.6 \times \frac{2}{8} = \frac{3}{20}$ | $0.4 \times \frac{3}{4} = \frac{3}{10}$ | $\frac{9}{20}$ |
| $1$ | $0.6$ | $0.4$ | $1$ |

Obtain joint probability: $P(X, Y)$

# Probability Theory (Rules of Probability)

Extension to continuous variables

**sum rule** $p(y) = \int p(x,y)\mathrm{d}x$

**product rule** $p(x,y) = p(x|y)p(y)$

$$p(y) = \int p(x,y)\mathrm{d}x$$

$$= \int p(y|x)p(x)\mathrm{d}x$$

We need to get familiar with this transformation process.

**Bayes' theorem** $p(y|x) = \dfrac{p(x|y)p(x)}{p(x)}$

when $x$ and $y$ are **independent**,

$$p(y|x) = p(x)$$

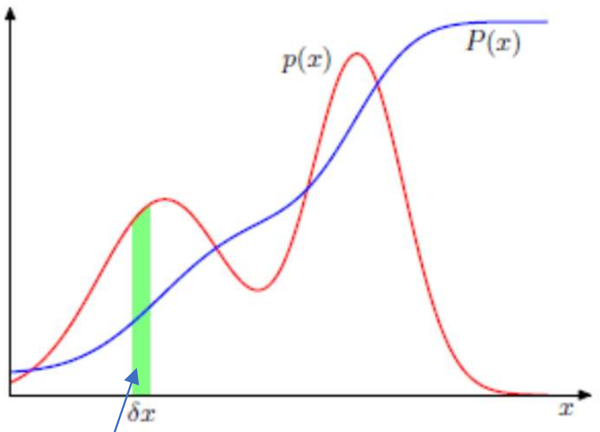
Therefore,

$$p(x,y) = p(x)p(y)$$

Please confirm this by following the rules of probability.

$$p(y|x) = \int p(y|g)p(g|x)\,\mathrm{d}g$$

# Probability Theory (Rules of Probability)




PRML, p. 18

$p(x)\delta x$: a probability

Required two conditions

$$\int_{-\infty}^{\infty} p(x)\mathrm{d}x = 1$$

$p(x) \geq 0$ $p(x)$ can be more than 1.

$p(x)$: **probability density function (pdf)**

$p(x)$ is not a probability.

$P(x)$: cumulative distribution function (cdf)

$$P(z) = \int_{-\infty}^{z} p(x)\mathrm{d}x$$

# Probability Theory

## Expectation (Mean) $\mu$

$$E[f] = \int f(x)p(x)\mathrm{d}x$$

$$\approx \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

consider the random variable $x$ itself

$$E[x] = \int xp(x)\mathrm{d}x$$

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

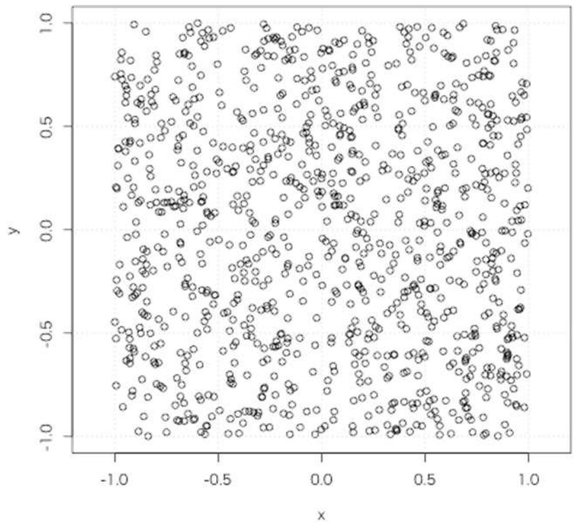
How to generate such points?

note:

**by sum and product rules**

$$p(y) = \int p(y|x)p(x)\mathrm{d}x$$

Monte Carlo (MC) sampling?

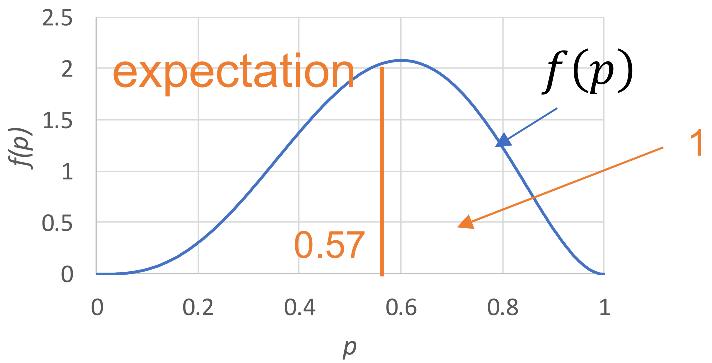



approximated by a finite number $N$ of points

- The points have to be generated according to the probability distribution $p(x)$.

# Probability Theory

Example: There is a pdf $f(p)$.




$$\int f(p)\mathrm{d}p = 1$$

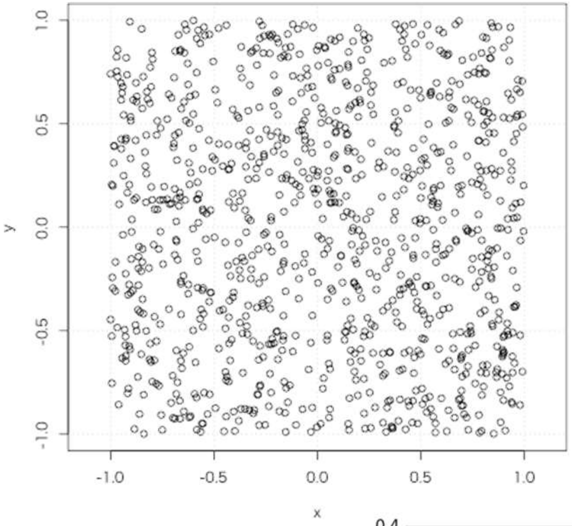
natural from the definition of pdf
meaning: the area under $f(p)$ is 1.

$$E[p] = \int p \times f(p)dp = 0.507$$

computing the expectation of the pdf $f(p)$
Meaning: the mean value of the variable $p$

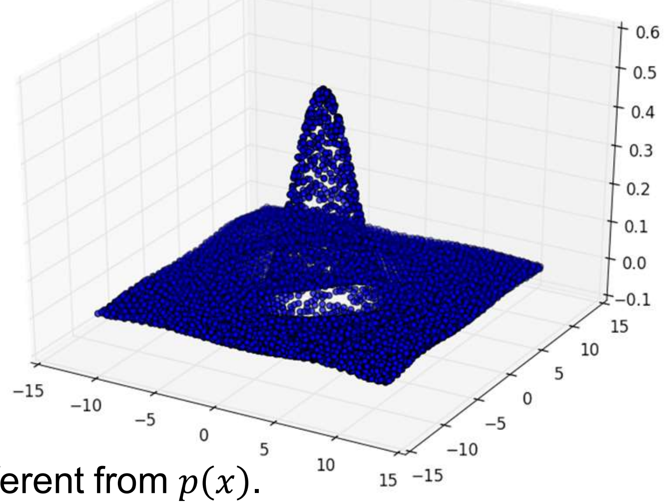
# Probability Theory

Monte Carlo sampling (random sampling)

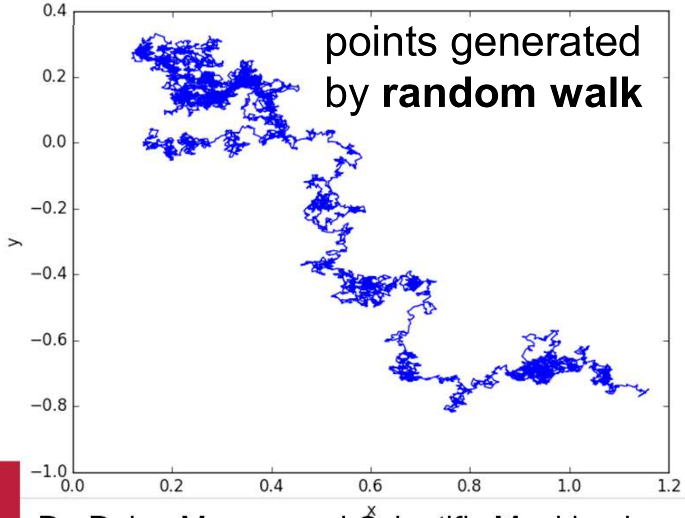


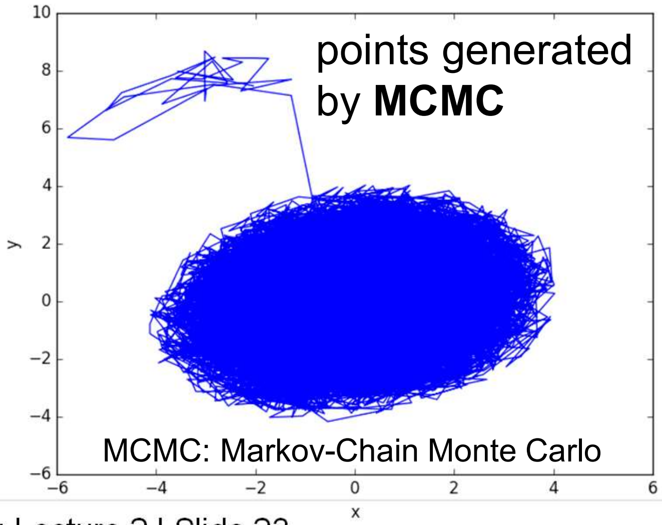
$p(x) = 1$
(**uniform distribution**)

visualization of the target pdf $p(x)$



Note:
$f(x)$ is another information different from $p(x)$.



# Probability Theory

consider random variable $x$

Variance $\sigma^2$

$var[x] = E[(x - E[x])^2]$ mean of the gap from the mean value of $f(x)$

useful property (not used in machine learning techniques)

$var[x] = E[x^2] - E[x]^2$

Standard deviation $\sigma$

$std[x] = \sqrt{var[x]}$ using the same unit as $x$

# Probability Theory

Covariance $\sigma_{x,y}$

(when standard deviation of random variables $x$ and $y$ are $\sigma_x$ and $\sigma_y$, respectively)

$$var[x] = E[(x - E[x])(x - E[x])] \qquad {\sigma_x}^2$$

$$cov[x, y] = E[(x - E[x])(y - E[y])] \qquad \sigma_{x,y}$$

Correlation: standardization of covariance

$$r_{x,y} = \frac{cov[x, y]}{\sigma_x \sigma_y} \qquad -1 \leq r_{x,y} \leq 1$$

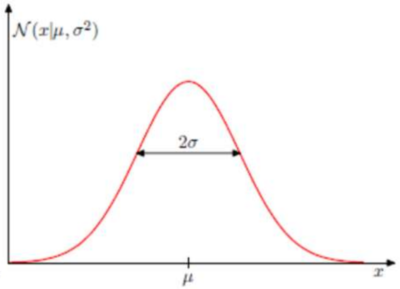
These indicators (covariance, correlation) do not always causal relationship.

The concept of covariance (correlation) is very important in various method in machine learning techniques.

# Probability Theory

Representation of a pdf $p(x)$

Gaussian distribution (as one example currently)




PRML, p. 25

The pdf $p(x)$ is **uniquely determined by two parameters**, $\mu$ and $\sigma$.

- Parametric distributions
  - Various distributions such as Gaussian distribution

- Non-parametric distributions
  - Distributions formed by sampling (the MCMC result in the previous slide)

$$\underbrace{p(x|\mu,\sigma)}_{\text{A pdf of } x \text{ when } \mu \text{ and } \sigma \text{ is given.}} = \mathcal{N}(x|\mu,\sigma^2)$$

We need to get familiar with this expression.

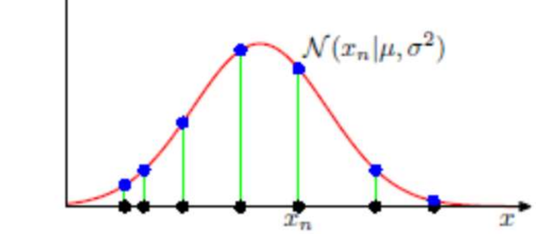
A pdf of $x$ when $\mu$ and $\sigma$ is given.

This rule can be used no matter whether $\mu$ and $\sigma$ are random variables or deterministic variables!

# Probability Theory

Likelihood function: a probability of data

$$\mathbf{X} = (x_1, x_2, \cdots, x_N)$$




Data points are assumed to be generated from <u>a distribution</u> (pdf) $p(x)(= p(x|\mu, \sigma))$.

1. **<u>Independent</u> and identically distributed (i.i.d.)**

$$p(x_1, x_2) = p(x_1)p(x_2) = \prod_{i=1}^{2} p(x_i)$$

2. $p(x_i|\mu, \sigma)$:
the probability when the data point $x_i$ is generated from the distribution $p(x|\mu, \sigma)$.

We can define the probability when all the data points are generated from the distribution $p(x|\mu, \sigma)$, which is $p(\mathbf{X}|\mu, \sigma)$.

**a probability of the data $\mathbf{X}$**

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

When this probability is regarded as a function of the parameters $\mu$ and $\sigma$, $p(\mathbf{X}|\mu, \sigma)$ is not a probability anymore.

But useful for estimation of the parameters $\mu, \sigma$!