# Mathematic Foundation and Basic Concept

Hao Dong

Peking University

# Recap: Prerequisites

- Basic knowledge of probabilities
  - Bayes rule, chain rule, probability distribution …
- Basic knowledge of information theory
  - Self-information, Shannon entropy, differential entropy …
  - Kullback-Leibler (KL) divergence
- Basic knowledge of machine learning/deep learning
  - “Machine Learning”, “Pattern Recognition and Machine Learning”
  - “Computer Vision”, “Natural Language Processing” …
- Basic programming language
  - Python

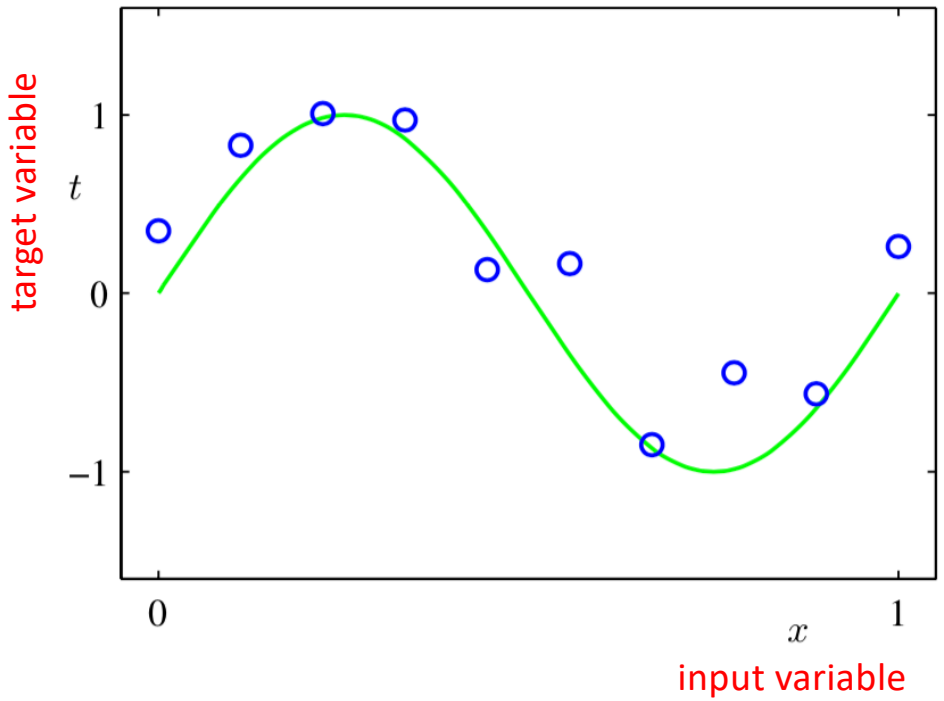
# Mathematic Foundation and Basic Concept

- Example: Regression - Polynomial Curve Fitting
- Probability Theory
- Decision Theory
- Information Theory

- Example: Regression - Polynomial Curve Fitting
- Probability Theory
- Decision Theory
- Information Theory

# Example: Polynomial Curve Fitting

- Problem Definition



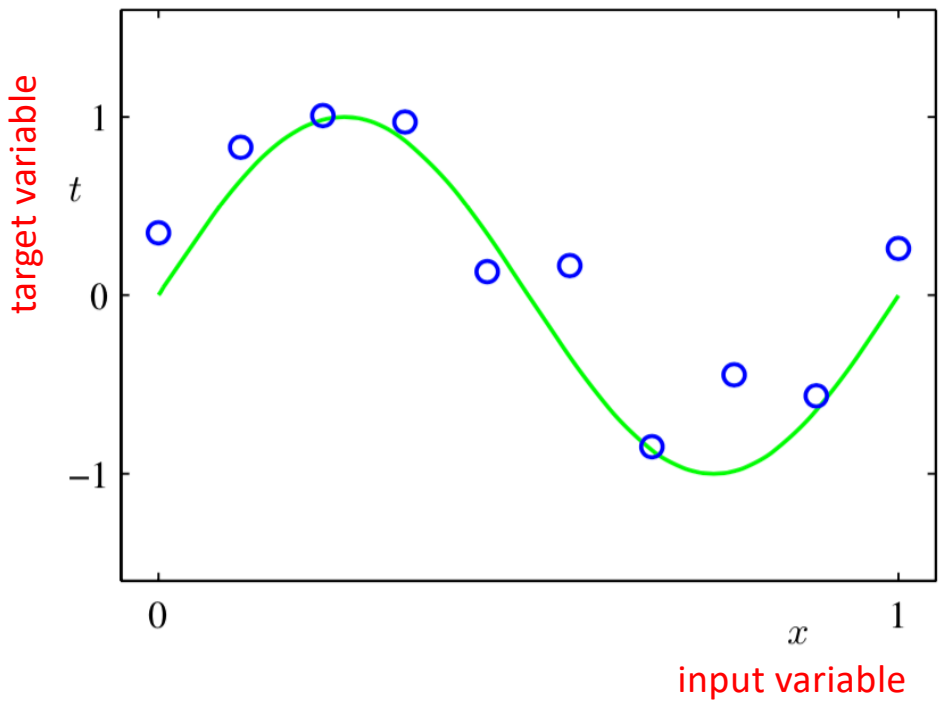

- A training dataset of $N = 10$ points
- Predict the target value $t$ given the input $x$

$\mathbf{x} \equiv (x_1, \ldots, x_N)^{\mathrm{T}}$
$\mathbf{t} \equiv (t_1, \ldots, t_N)^{\mathrm{T}}$

A row vector with $N$ elements

# Example: Polynomial Curve Fitting

- Simple Model: Polynomial Function




$M$ order polynomial:

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

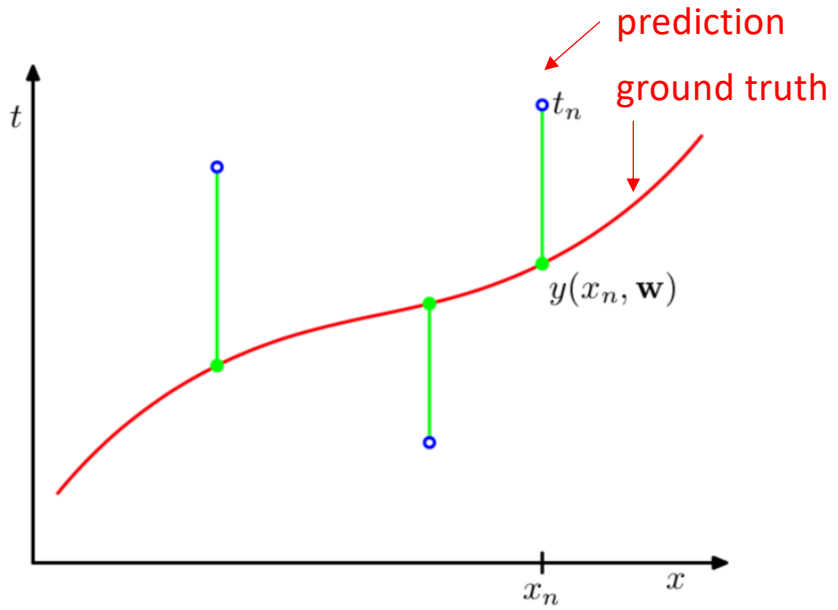
$\mathbf{w} = (w_1, \ldots, w_M)^{\mathrm{T}}$ Weights: a row vector with $M$ elements

The appropriate value $\hat{t}$:

$$\hat{t} = \mathbf{x}^{\mathrm{T}}\mathbf{w}$$

# Example: Polynomial Curve Fitting

- Error Function




Minimize an error function

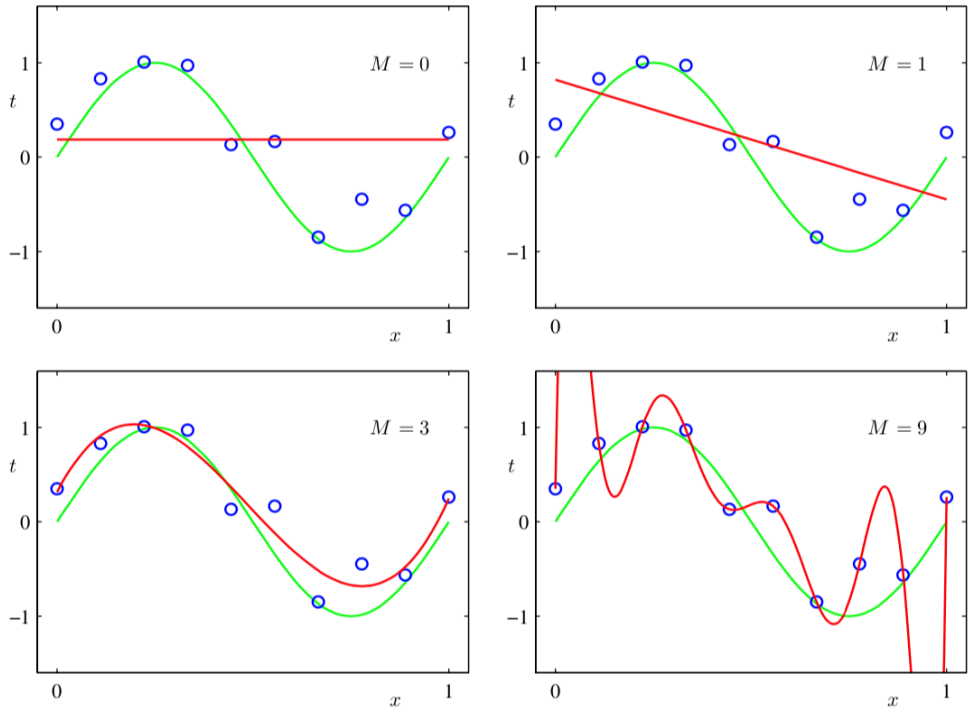
$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$$

Optimize the weights

"$\frac{1}{2}$" is for later convenience, we will discuss later

# Example: Polynomial Curve Fitting

- Model Capacity and Overfitting




$M$ order polynomial

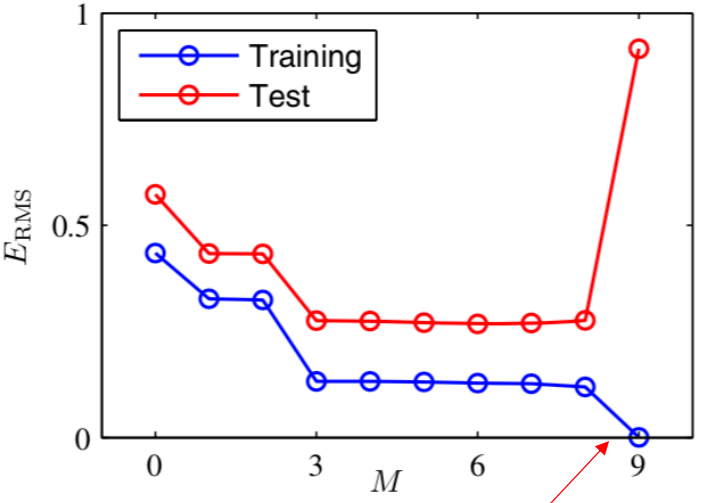
Root mean square error:

$$E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^\star)/N}$$

Averaged/mean error

$$E(\mathbf{w}^\star) = 0$$

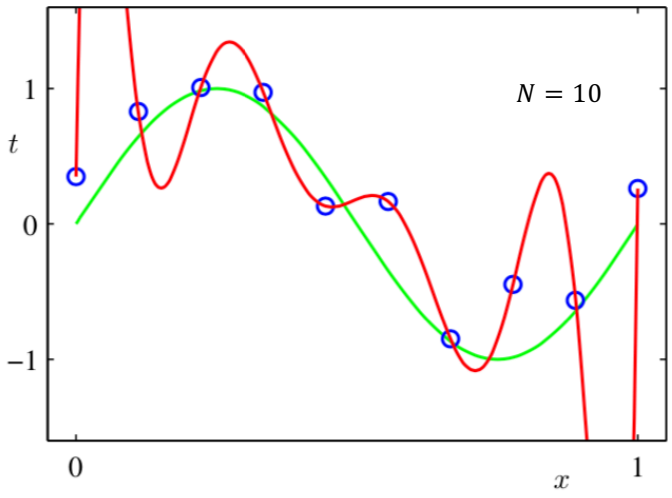
Optimized weights

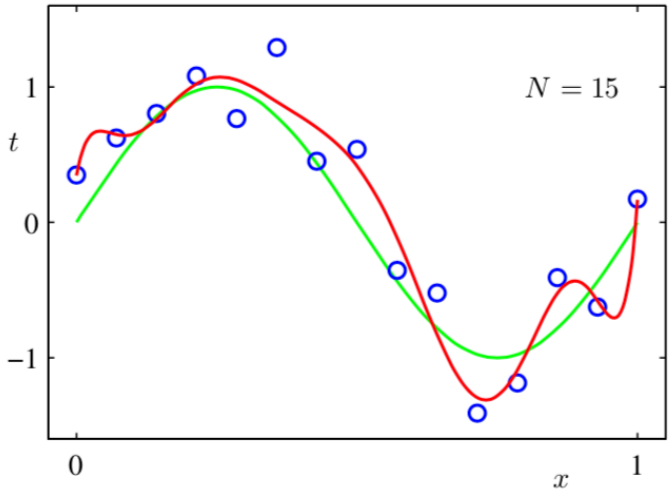


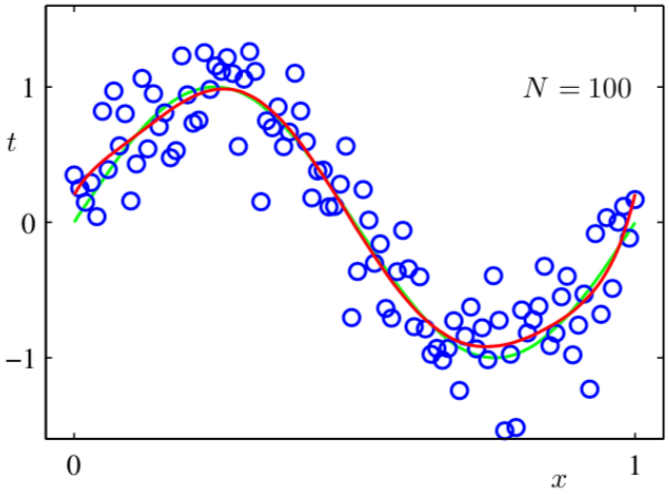

Large $M$ will lead to overfitting

# Example: Polynomial Curve Fitting

- Solution1 : More Training Data Point




$M = 9$ with $N = 10, 15, 100$

More data points = Better generalization

# Example: Polynomial Curve Fitting

- Solution2 : Weight Regularization

| | $M=0$ | $M=1$ | $M=6$ | $M=9$ |
|---|---|---|---|---|
| $w_0^\star$ | 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 |

Very large

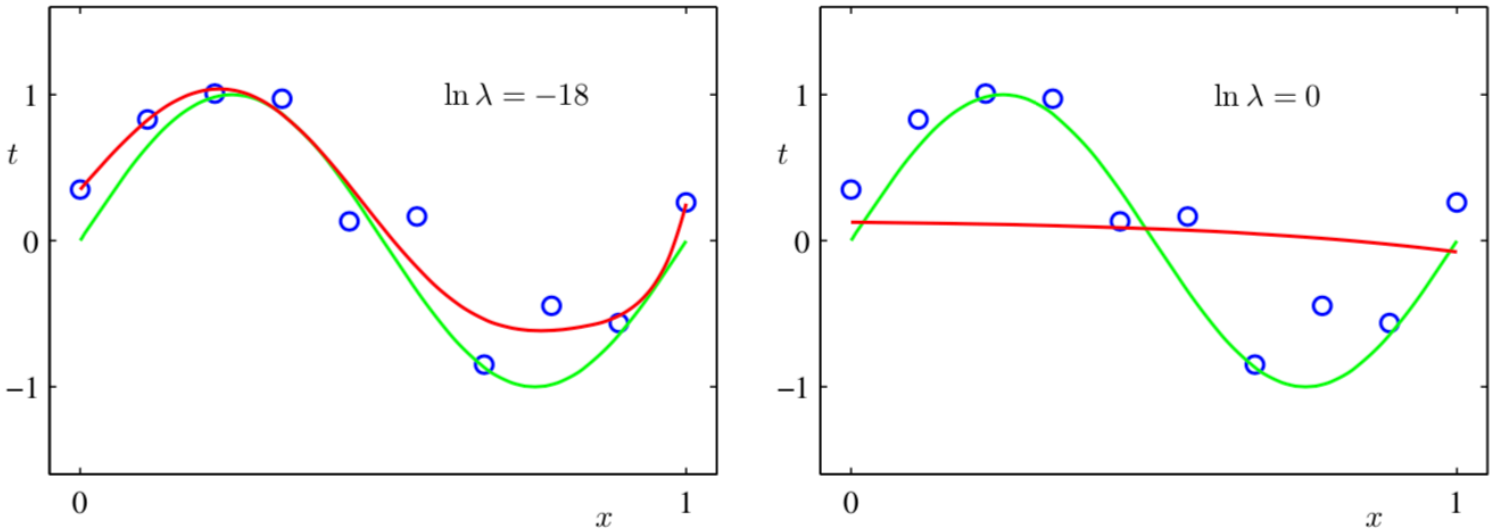
Optimized weights

# Example: Polynomial Curve Fitting

- Solution2 : Weight Regularization

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

$$\|\mathbf{w}\|^2 \equiv \mathbf{w}^{\mathrm{T}}\mathbf{w} = w_0^2 + w_1^2 + \ldots + w_M^2$$

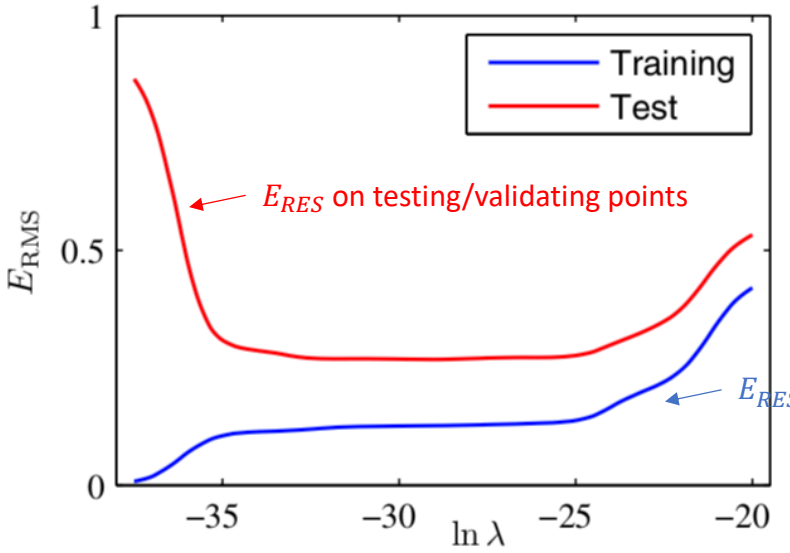



$M = 9$, with different $\lambda$

Larger $\lambda$ = Stronger regularization

# Example: Polynomial Curve Fitting

- Solution2 : Weight Regularization




$M = 9$, with different $\lambda$

Larger $\lambda$ = Stronger regularization

- Example: Regression - Polynomial Curve Fitting
- Probability Theory
- Decision Theory
- Information Theory

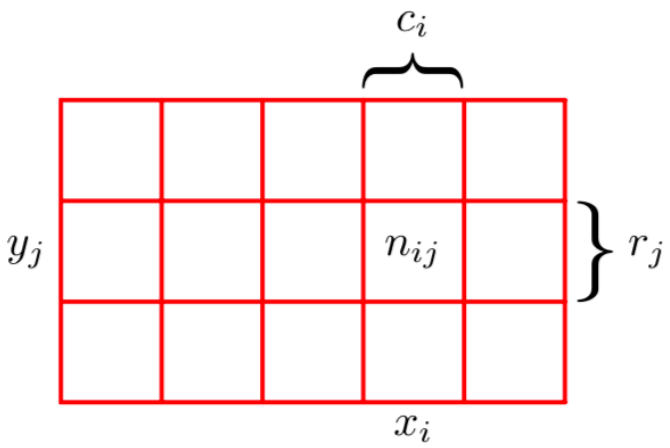
# Probability Theory

- The probability of an event is the fraction of times that event occurs out of the total number of trials
- The probability must lie in the interval [0, 1]

- A box contains red and blue balls, we randomly pick 100 balls from a box, and 90 balls are red

$$p(ball = red) = \frac{90}{100} = 0.9$$

# Probability Theory




**Joint probability:** $X = x_i$ and $Y = y_i$ happen together

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

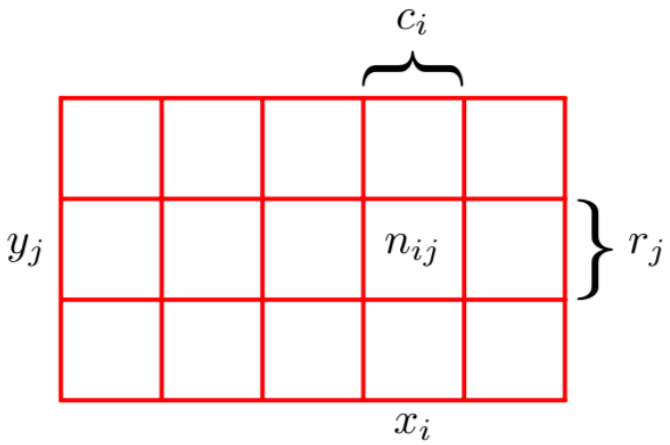
**Given ..**

$$p(X = x_i) = \frac{c_i}{N}$$

So.. We have the **sum rule of probability**

$$p(X = x_i) = \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

# Probability Theory




Variable Value

**Conditional probability:** $Y = y_i$ given $X = x_i$

$$p(Y = y_j|X = x_i) = \frac{n_{ij}}{c_i}$$

As we know:

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} \qquad p(X = x_i) = \frac{c_i}{N}$$

We can have the **product rule of probability**

$$\begin{aligned} p(X = x_i, Y = y_j) &= \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} \\ &= p(Y = y_j|X = x_i)p(X = x_i) \end{aligned}$$

# Probability Theory

- The Rules of Probability

$$\textbf{sum rule} \qquad p(X) = \sum_{Y} p(X, Y)$$

$$\textbf{product rule} \qquad p(X, Y) = p(Y|X)p(X)$$

# Probability Theory

- Bayes' theorem

Given the **product rule**:

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

symmetry property $p(X,Y) = p(Y,X)$

We can have the **Bayes' theorem:**

$$p(Y|X) = \frac{p(X,Y)}{p(X)} = \frac{p(X|Y)p(Y)}{p(X)}$$

Using the **sum rule (** $p(X) = \sum_Y p(X,Y)$ **)**, we can have:

$$p(X) = \sum_Y p(X|Y)p(Y)$$

# Probability Theory

- Bayes' theorem

**The prior probability:** it is the probability available before we observe the $X$

先验

**The likelihood：** It expresses how probable the observed data set is for different settings of the parameter y

似然

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

**The normalization constant** required to ensure that the sum of the conditional probability over all values of Y equals one.

$$p(X) = \int_Y p(X|Y)p(Y)dY$$

**The posterior probability:** it is the probability obtained after we have observed $X$

后验

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

# Probability Theory

- The Rules of Probability

**sum rule** $$p(X) = \sum_{Y} p(X, Y)$$

**product rule** $$p(X, Y) = p(Y|X)p(X)$$

**Bayes' theorem** $$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

where $$p(X) = \sum_{Y} p(X|Y)p(Y)$$

# Probability Theory

- Independent Events

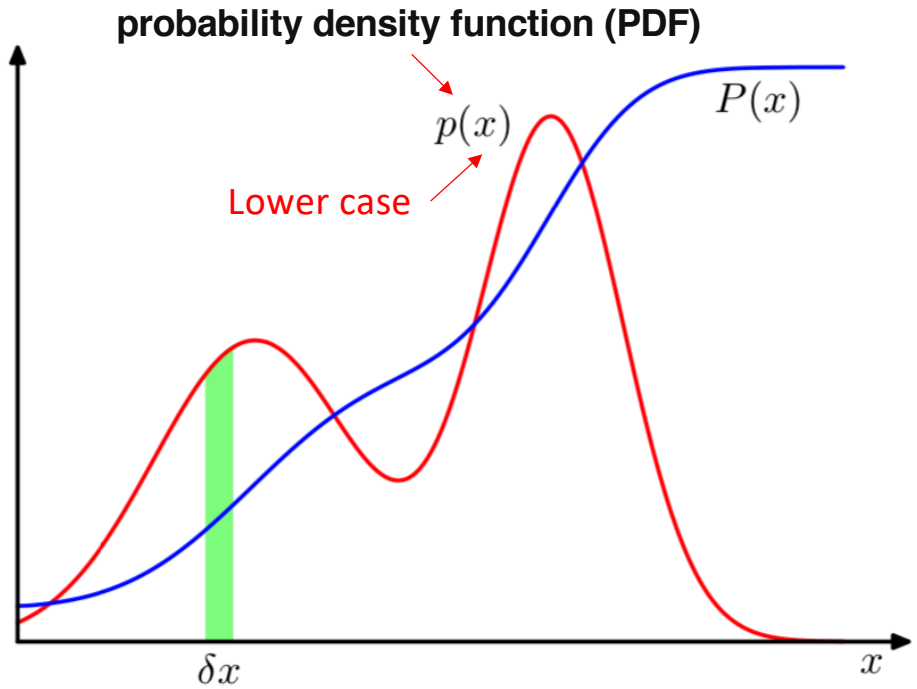
**If $X$ and $Y$ are independent**

**Bayes' theorem** $$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$ $$p(Y|X) = p(Y)$$

**product rule** $$p(X,Y) = p(Y|X)p(X)$$ $$p(X,Y) = p(X)p(Y)$$

# Probability Theory

- Probability Densities




The probability that $x$ will lie in an interval $(a, b)$ is given by:

$$p(x \in (a,b)) = \int_a^b p(x)\,\mathrm{d}x$$

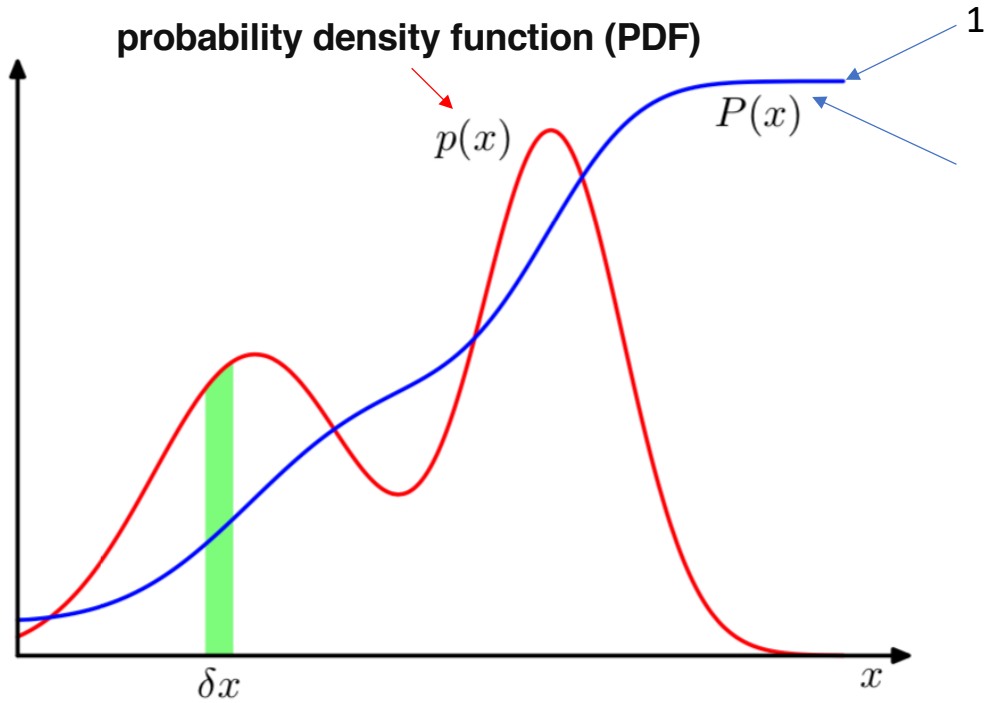
Also ..

$$\begin{aligned} p(x) &\geqslant 0 \\ \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x &= 1 \end{aligned}$$

If the probability of a real-valued variable $x$ falling in the interval $(x, x + \delta x)$ is given by $p(x)\delta x$ for $\delta x \to 0$, then $p(x)$ is called **the probability density** over x.

# Probability Theory

- Probability Densities




**cumulative distribution function**

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

which satisfies $P'(x) = p(x)$

# Probability Theory

- **Expectations** and Covariances

The average value of some function $f(x)$ under a probability distribution $p(x)$ is called the **expectation** of $f(x)$ and will be denoted by $\mathrm{E}[f]$:

For a continuous distribution:

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

For a discrete distribution:

$$\mathbb{E}[f] = \sum_{x} p(x) f(x)$$

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^{N} f(x_n)$$

It becomes exact in the limit $N \to \infty$

# Probability Theory

- **Expectations** and Covariances

In the case of **continuous variables**, expectations are expressed in terms of an integration with respect to the corresponding probability density

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

**The conditional expectation** with respect to a conditional distribution

$$\mathbb{E}_x[f|y] = \sum_x p(x|y) f(x)$$

# Probability Theory

- Expectations and **Covariances**

The **variance** of $f(x)$:

$$\begin{aligned}\mathrm{var}[f] &= \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right] \\ &= \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2\end{aligned}$$

The **variance** of the variable $x$:

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$$

For two random variables $x$ and $y$, the **covariance** is:

$$\begin{aligned}\mathrm{cov}[x, y] &= \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]\end{aligned}$$

For two vectors of random variables **x** and **y**, the **covariance matrix:**

$$\begin{aligned}\mathrm{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\{\mathbf{y}^{\mathrm{T}} - \mathbb{E}[\mathbf{y}^{\mathrm{T}}]\}\right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}].\end{aligned}$$

The covariance of the components of a vector **x** with each other:

$$\mathrm{cov}[\mathbf{x}] \equiv \mathrm{cov}[\mathbf{x}, \mathbf{x}]$$

# Probability Theory

- Bayesian Probabilities

  - **Frequentist**: Probability in terms of frequencies of random, repeatable events.
    - Do not work if the event is not repeatable, e.g., the probability of the Arctic ice cap disappear.

  - **Bayesian**: Degree of belief.

# Probability Theory

- Frequentist Probability

  - Maximum Likelihood

A widely used frequentist estimator is maximum likelihood, in which $\theta$ is set to the value that maximizes the likelihood function p(X| $\theta$).

- Given $x \sim \text{iid} \sim p(x|\theta)$ , $p(X|\theta) = \prod_{i=1}^{N} p(x_i|\theta)$
- $\theta_{\text{MLE}} = arg\max_{\theta} logp(X|\theta) = arg\max_{\theta} \sum_{i=1}^{N} logp(x_i|\theta)$

# Probability Theory

- Bayesian Probabilities

  - $\theta$: random variables, $\theta \sim p(\theta)$ ← prior
  - MAP: Maximum A Posterior

likelihood prior

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)} \propto p(X|\theta)p(\theta)$$

posterior normalization constant

  - $\theta_{\text{MAP}} = arg\max_{\theta} p(\theta|X) = arg\max_{\theta} p(X|\theta)p(\theta)$

Bayes can evaluate the **uncertainty** in $\theta$ *after* we have observed $X$ in the form of the posterior probability $p(\theta|X)$.

# Probability Theory

- Bayesian Probabilities
  - $\theta$: random variables, $\theta \sim p(\theta)$ ← prior
  - Bayesian Estimation

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)} = \frac{p(X|\theta)p(\theta)}{\int_\theta p(X|\theta)p(\theta)d\theta}$$

  - Bayesian Prediction
    - Given dataset $X$ and a sample $\tilde{x}$ find $p(\tilde{x}|X)$ via $\theta$

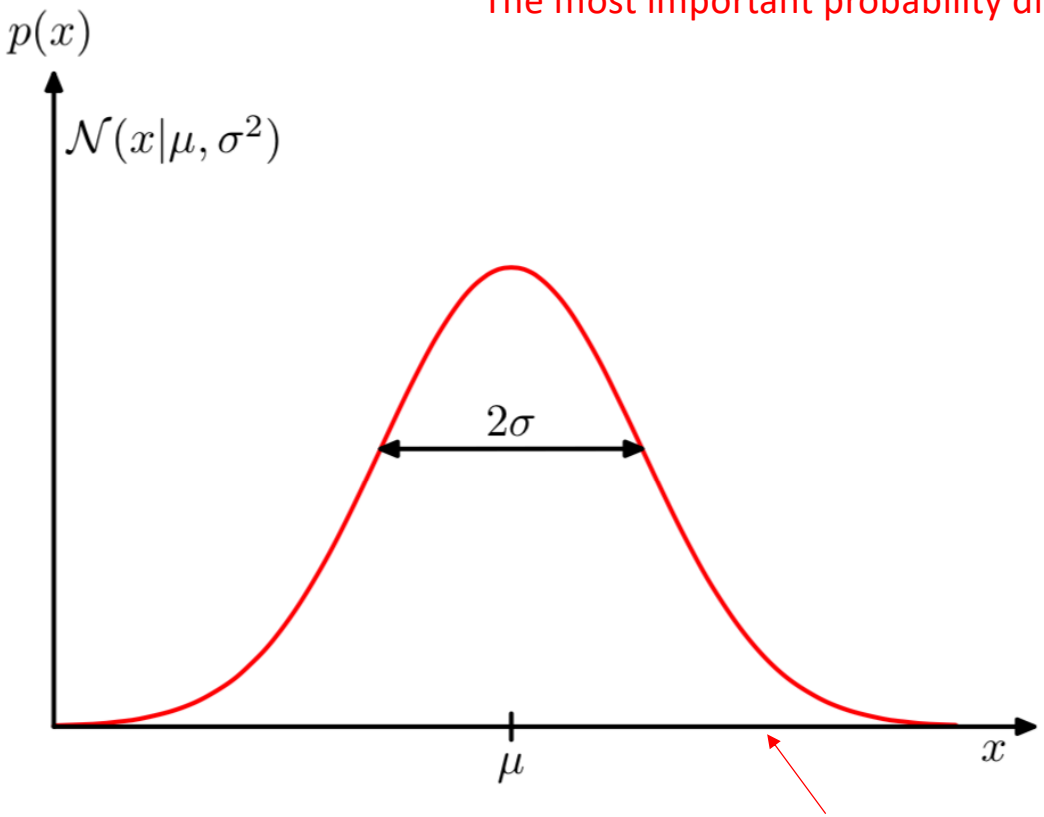
$$p(\tilde{x}|X) = \int_\theta p(\tilde{x}, \theta|X)d\theta = \int_\theta p(\tilde{x}|\theta)p(\theta|X)d\theta =$$

posterior

# Probability Theory

- The Gaussian Distribution

The most important probability distribution for continuous variables




$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

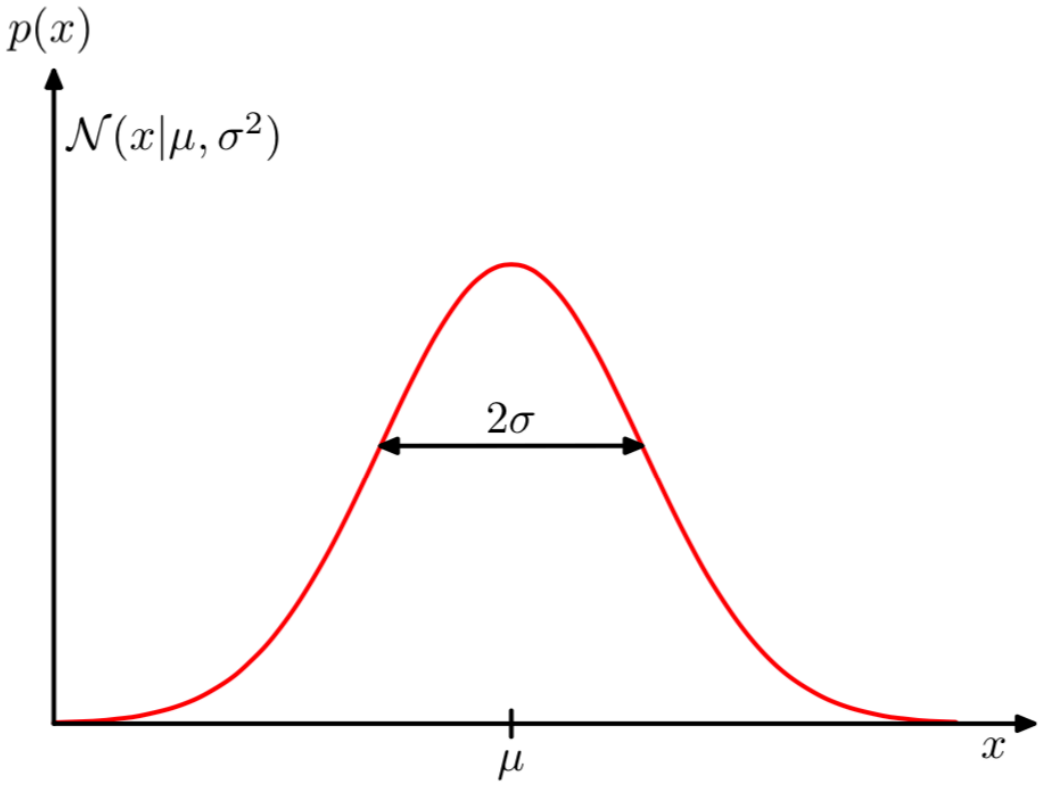
$$\mathcal{N}(x|\mu,\sigma^2) > 0 \qquad \int_{-\infty}^{\infty}\mathcal{N}\left(x|\mu,\sigma^2\right)\,\mathrm{d}x = 1$$

$\mu$: mean
$\sigma$: standard deviation
$\sigma^2$: variance
$\beta = \frac{1}{\sigma^2}$: precision

Also called Normal distribution when mean=0 and variance=1

# Probability Theory

- The Gaussian Distribution




$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) x \, \mathrm{d}x = \mu$$

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) x^2 \, \mathrm{d}x = \mu^2 + \sigma^2$$

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

# Probability Theory

- The Gaussian Distribution

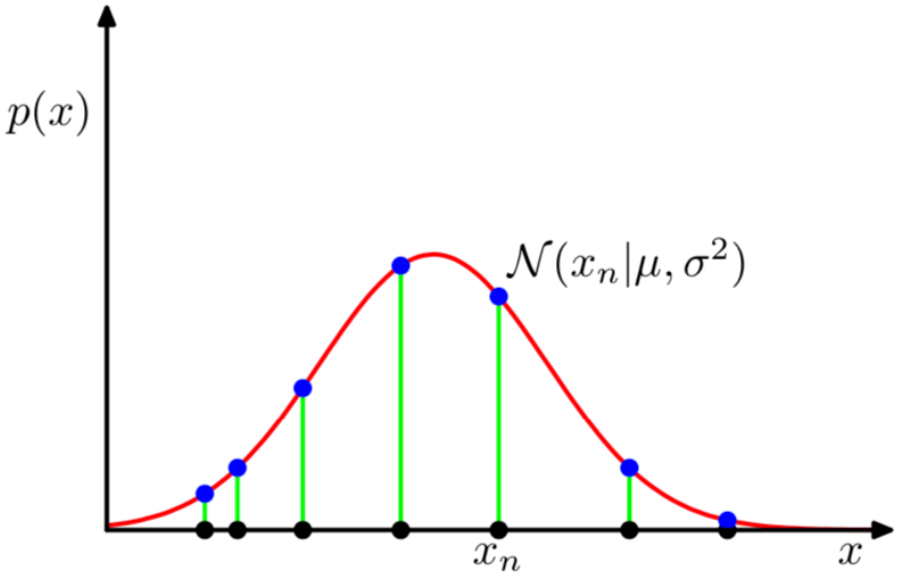
**D-dimensional vector x of continuous variables**

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

$\mu$ is called the mean, the D × D matrix
Σ is called the covariance
|Σ| denotes the determinant of Σ

# Probability Theory

- The Gaussian Distribution




Now suppose that we have a dataset of observations $\mathbf{x} = (x_1, \dots, x_N)^{\mathrm{T}}$, representing $N$ observations of the scalar variable $x$.

How to determine the mean and variance according to the dataset??

All data points are sampled independently from the same distribution, they are **independent and identically distributed (i.i.d).**

For i.i.d variables, we can have the **joint probability** as follow

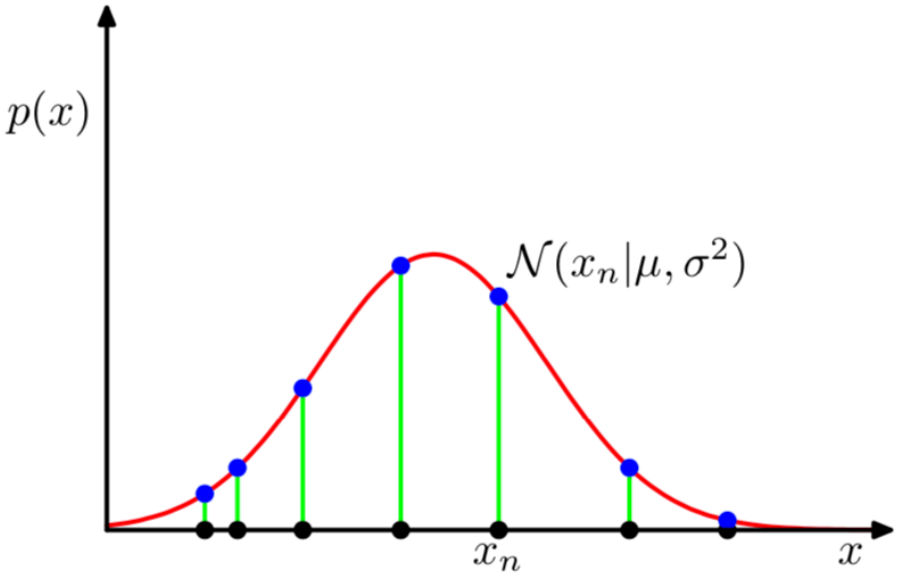
$$p(\mathbf{x}) = p(x_1)\, p(x_1)\, p(x_1) \dots .\, p(x_N)$$

So…

$$p(\mathbf{x}|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(x_n|\mu,\sigma^2\right)$$

# Probability Theory

- The Gaussian Distribution




**Maximum Likelihood**

To find the "best" mean and variance, we can maximize the probability of the parameters given the data

$$p(\mathbf{x}|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(x_n|\mu,\sigma^2\right)$$

In practice, it is more convenient to maximize the log of the likelihood function. The logarithm is a monotonically increasing function, maximization of the log of a function is equivalent to maximization of the function itself

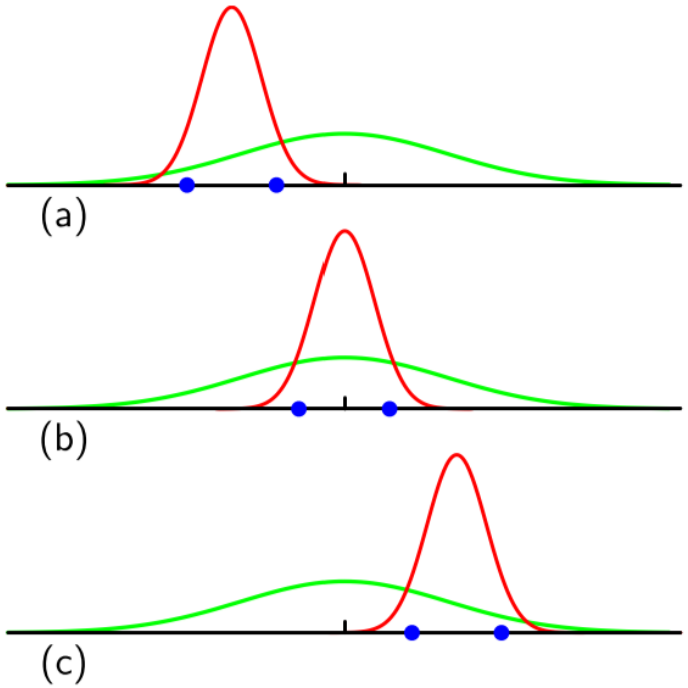
$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

So.. the **maximum likelihood solution**:

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n \qquad \sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2$$

# Probability Theory

- The Gaussian Distribution




$$\mathbb{E}[\mu_{\mathrm{ML}}] = \mu$$
$$\mathbb{E}[\sigma^2_{\mathrm{ML}}] = \left(\frac{N-1}{N}\right)\sigma^2$$

On average the maximum likelihood estimate will obtain the correct mean but will underestimate the true variance by a factor (N − 1)/N.
The following estimate for the variance parameter is unbiased.

$$\widetilde{\sigma}^2 = \frac{N}{N-1}\sigma^2_{\mathrm{ML}} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2$$

- Example: Regression - Polynomial Curve Fitting
- Probability Theory
- Decision Theory
- Information Theory

# Decision Theory

- An example
- Basic intuition
- Minimizing the misclassification rate
- Minimizing the expected loss
- The reject option
- Loss functions for regression

# Decision Theory

- A example

A medical diagnosis problem in which we have taken an X-ray image of a patient, and we wish to determine whether the patient has cancer or not.

We are interested in **the probabilities of the two classes** given the **image**, which are given by p(Ck|x). Using Bayes' theorem, these probabilities can be expressed in the form :

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})}.$$

Note that any of the quantities appearing in Bayes' theorem can be obtained from the joint distribution p(x, Ck) by either marginalizing or conditioning with respect to the appropriate variables.

We can now interpret **p(Ck)** as the prior probability for the class Ck, and p(Ck|x) as the corresponding posterior probability. Thus p(C1) represents the probability that a person has cancer, before we take the X-ray measurement.

Similarly, p(C1|x) is the corresponding probability, revised using Bayes' theorem in light of the information contained in the X-ray.

# Decision Theory

- Basic intuition

If our aim is to minimize the chance of assigning x to the wrong class, then intuitively we would choose the class having the higher posterior probability. We now show that this intuition is correct, and we also discuss more general criteria for making decisions.

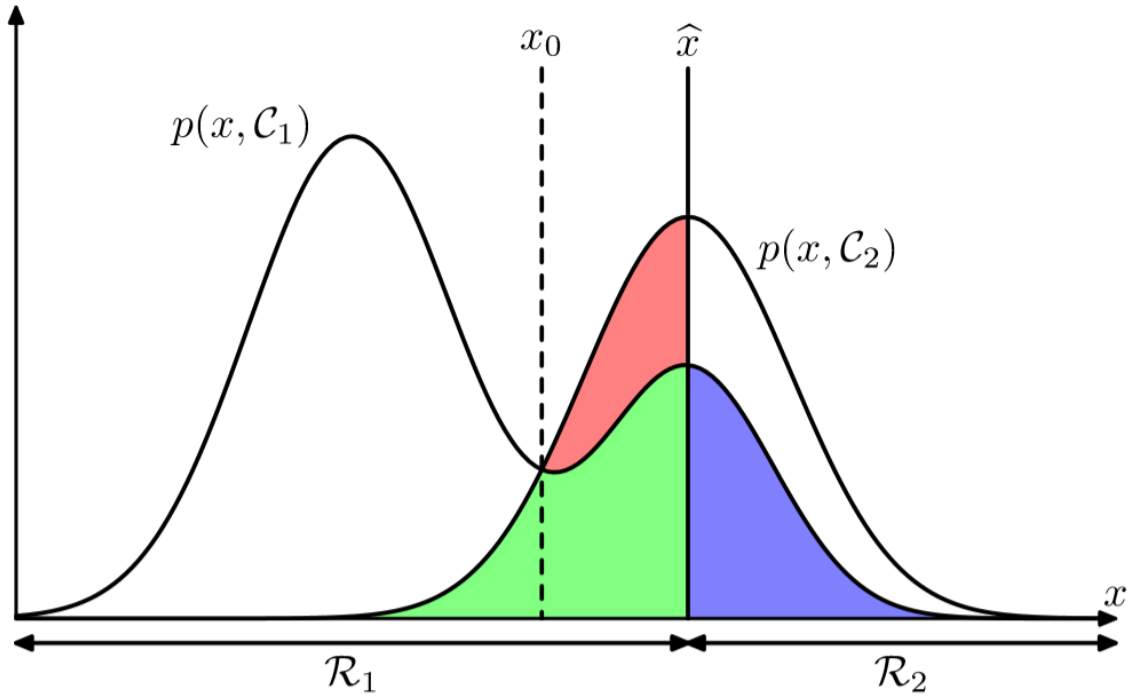
# Decision Theory

- Minimizing the misclassification rate

A mistake occurs when an input vector belonging to class C1 is assigned to class C2 or vice versa. The probability of this occurring is given by

$$\begin{aligned} p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\, \mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\, \mathrm{d}\mathbf{x}. \end{aligned}$$

# Decision Theory



# Decision Theory

- Minimizing the expected loss

Loss function, also called a cost function, which is a single, overall measure of loss incurred in taking any of the available decisions or actions.

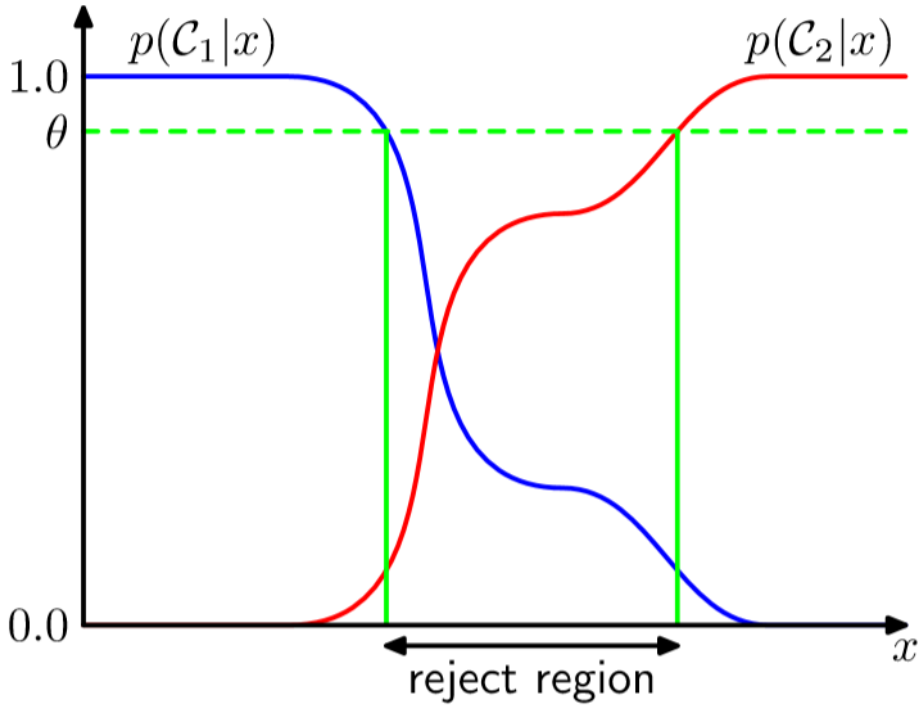
Our goal is then to minimize the total loss incurred.

We seek to minimize the average loss, where the average is computed with respect to this distribution, which is given by

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) \, \mathrm{d}\mathbf{x}$$

# Decision Theory

- The reject option



# Decision Theory

- Loss functions for regression

For regression problems, loss is given by

$$\mathbb{E}[L] = \iint L(t, y(\mathbf{x}))p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t.$$

A common choice of loss function in regression problems is the squared loss. In this case, the expected loss can be written:

$$\mathbb{E}[L] = \iint \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t.$$

We can expand the square term as follows:

$$\begin{aligned}\{y(\mathbf{x}) - t\}^2 &= \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}] + \mathbb{E}[t|\mathbf{x}] - t\}^2 \\ &= \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 + 2\{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}\{\mathbb{E}[t|\mathbf{x}] - t\} + \{\mathbb{E}[t|\mathbf{x}] - t\}^2\end{aligned}$$

Expectation = 0

We obtain an expression for the loss function in the form:

$$\mathbb{E}[L] = \int \{y(\boldsymbol{x}) - \mathbb{E}[t \mid \boldsymbol{x}]\}^2 p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + \int \mathrm{var}[t \mid \boldsymbol{x}] p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

- Example: Regression - Polynomial Curve Fitting
- Probability Theory
- Decision Theory
- Information Theory

## Basic intuition

- Learning that an unlikely event has occurred is more informative than learning that a likely event has occurred.
  - E.g.
    - “the sun rose this morning” : uninformative
    - “there was a solar eclipse this morning”: informative.

## Basic intuition

- We would like to quantify information in a way that formalizes this intuition. Specifically:
  - Likely events should have low information content, and in the extreme case, events that are guaranteed to happen should have no information content whatsoever.
  - Less likely events should have higher information content.
  - Independent events should have additive information.
    - E.g. Finding out that a tossed coin has come up as heads twice should convey twice as much information as finding out that a tossed coin has come up as heads once.

## Self-information

- In order to satisfy all three of these properties, we define the self information of an event x to be

$$I(x) = -\log P(x).$$

(use log to mean the natural logarithm, with base e. I(x) is therefore written in units of nats)

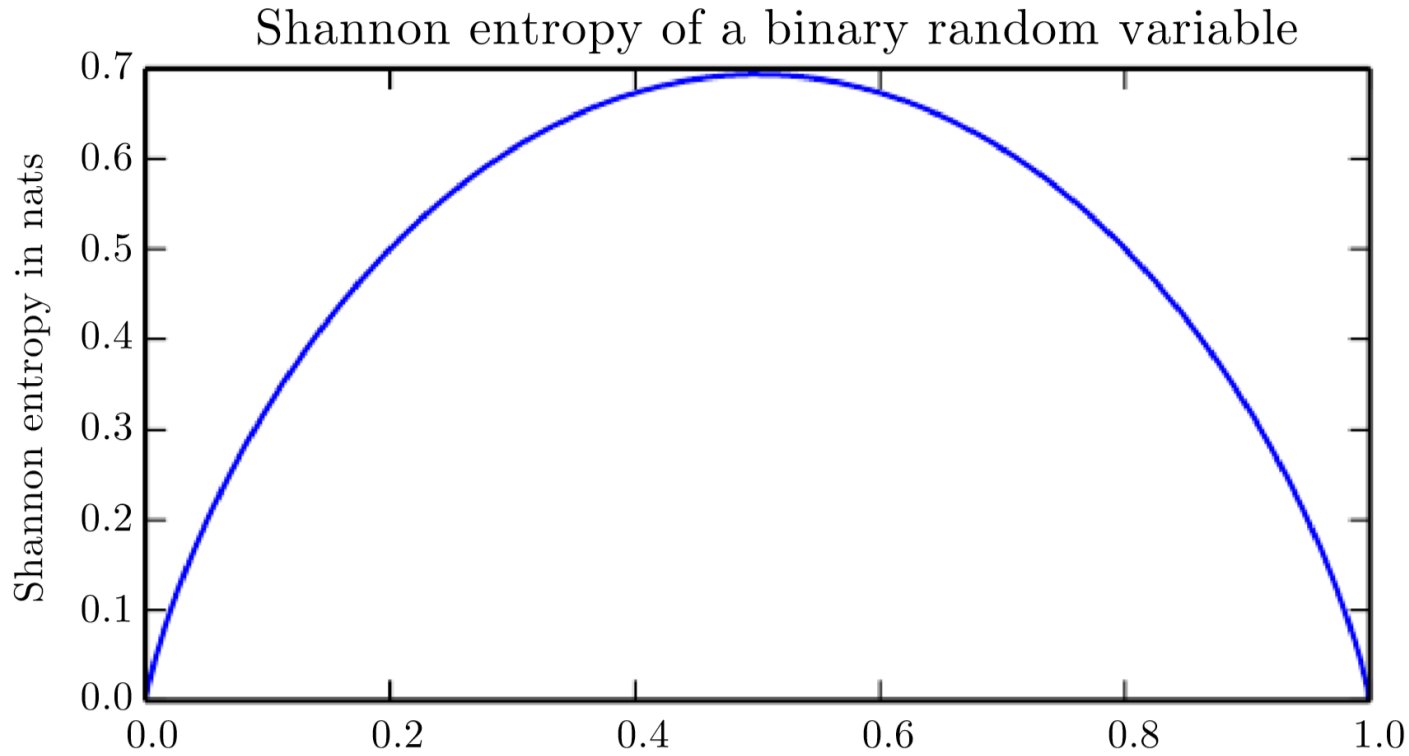
## Shannon entropy

- Self-information deals only with a single outcome. We can quantify the amount of uncertainty in an entire probability distribution using the Shannon entropy:

$$H(\mathrm{x}) = \mathbb{E}_{\mathrm{x}\sim P}[I(x)] = -\mathbb{E}_{\mathrm{x}\sim P}[\log P(x)].$$

  - The Shannon entropy of a distribution is the expected amount of information in an event drawn from that distribution. It gives a lower bound on the number of bits (if the logarithm is base 2, otherwise the units are different) needed on average to encode symbols drawn from a distribution P.

# Example for Shannon entropy



## KL divergence

- If we have two separate probability distributions P(x) and Q(x) over the same random variable x, we can measure how different these two distributions are using the Kullback-Leibler (KL) divergence:

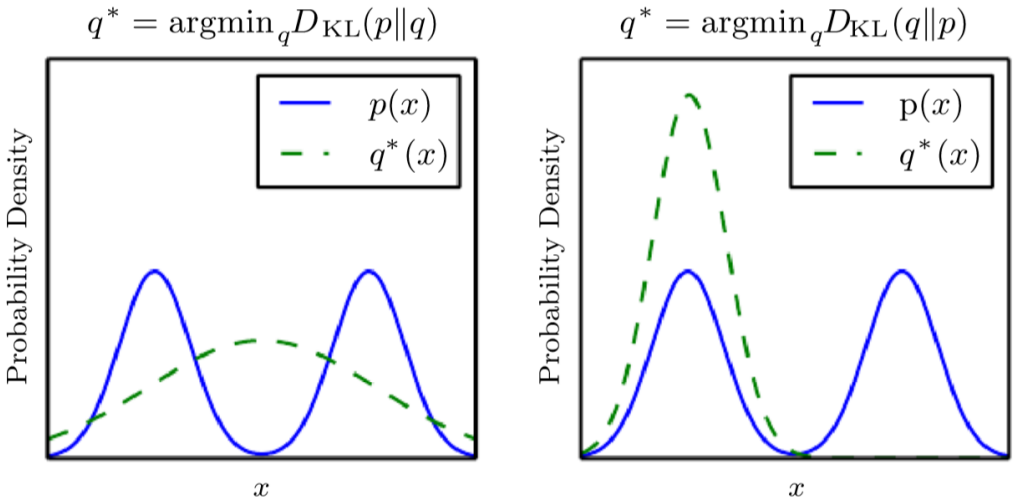
$$D_{\mathrm{KL}}(P\|Q) = \mathbb{E}_{\mathrm{x}\sim P}\left[\log \frac{P(x)}{Q(x)}\right] = \mathbb{E}_{\mathrm{x}\sim P}\left[\log P(x) - \log Q(x)\right].$$

  - In the case of discrete variables, it is the extra amount of information needed to send a message containing symbols drawn from probability distribution P, when we use a code that was designed to minimize the length of messages drawn from probability distribution.

## KL divergence

- Because the KL divergence is non-negative and measures the difference between two distributions, it is often conceptualized as measuring some sort of distance between these distributions.
  - Not a true distance measure because it is not symmetric.
  - It is nonnegative.

# KL divergence




- Which direction of the KL divergence to use?
  - Some applications require an approximation that usually places high probability anywhere that the true distribution places high probability: left one
  - Other applications require an approximation that rarely places high probability anywhere that the true distribution places low probability: right one

$$D_{\text{KL}}(P\|Q) = \mathbb{E}_{\text{x}\sim P}\left[\log \frac{P(x)}{Q(x)}\right] = \mathbb{E}_{\text{x}\sim P}\left[\log P(x) - \log Q(x)\right].$$

## Cross-entropy

- A quantity that is closely related to the KL divergence is the cross-entropy H(P,Q) = H(P) + $D_{KL}$(P||Q), which is similar to the KL divergence but lacking the term on the left:

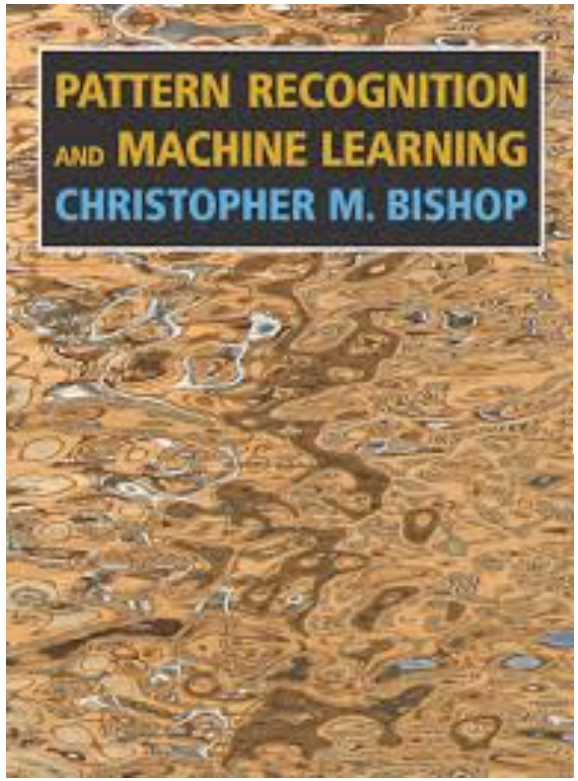
$$H(P, Q) = -\mathbb{E}_{\mathrm{x} \sim P} \log Q(x).$$

- Minimizing the cross-entropy with respect to Q is equivalent to minimizing the KL divergence, because does not participate in the omitted term.

# Mathematic Foundation and Basic Concept

- Example: Regression - Polynomial Curve Fitting
- Probability Theory
- Decision Theory
- Information Theory

# Reference




**Free Download**

https://github.com/zsdonghao/deep-learning-book/blob/master/All-in-one-pdf
《Deep Learning》all-in-one.pdf

# Thanks