Probability refresher

Advanced topics in AI-I (CK0146)

Francesco Corona

Outline

1. Probability theory
   - Probability densities
   - Expectations and covariances
   - Bayesian probabilities
   - The Gaussian distribution

2. Polynomial fitting
   - Polynomial fitting revisited
   - Bayesian polynomial fitting

3. Graphical models
   - Bayesian networks
   - Bayesian polynomial fitting

Probability theory

Probability theory

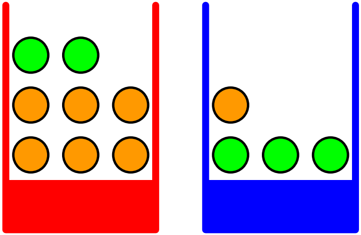
A key concept in the field of probabilistic modelling is that of **uncertainty**

- Gets in the way through noise on measurements
- Gets in the way through the finite size of data

Probability theory provides a consistent framework for the quantification and manipulation of uncertainty and forms one of the central foundations of PRML

Probability theory (cont.)

Imagine we have two boxes,one red and one blue, and in the red box we have 2 apples and 6 oranges, and in the blue box we have 3 apples and 1 orange



Suppose that we randomly pick one of the boxes and from that box we randomly select an item of fruit

- we check the fruit and we replace it in its box

We repeat this process *many* times

40% of the time we pick the red box and 60% of the time we pick the bluebox

- We are equally likely to select any of the pieces of fruit from the box

Probability theory (cont.)

The **identity of the box** that will be chosen is a **random variable** B

This random variable can take only two possible values

- either r, for red box or b, for blue box

The **identity of the fruit** that will be chosen is a **random variable** F

This random variable can take only two possible values

- either a, for apple or o, for orange

Probability theory (cont.)

Definition

We *define* the **probability of an event** to be the fraction of times that event occurs out of the total number of trials (*in the limit* that it goes to infinity)

Example

- The probability of selecting the red box is 4/10
- The probability of selecting the blue box is 6/10

We write these probabilities as $p(B = r) = 4/10$ and $p(B = b) = 6/10$

Probability theory (cont.)

Note that, by definition, **probabilities must lie in the interval** $[0, 1]$

- If the events are **mutually exclusive** and if they **include all possible outcomes**, then the **probabilities** for those events **must sum to one**

Example

We have defined our experiment and we can start asking questions

- What is the overall probability that the selection procedure picks an apple?
- Given that we have chosen an orange, what is the probability that the box we chose was the blue one?
- ...

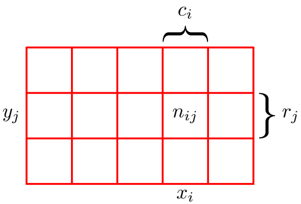
We can answer questions such as these, and indeed much more complex questions associated with problems in pattern recognition, once we have equipped ourselves with the **two elementary rules of probability**

- the **sum rule** and the **product rule**

Probability theory (cont.)

To derive the rules of probability, consider the slightly more general example

- **Two random variables** X and Y



We shall suppose that:

- **X can take any of the values x_i, $i = 1, \ldots, M$**
- **Y can take any of the values y_j, $j = 1, \ldots, L$**

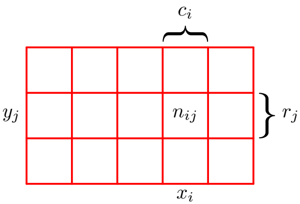
Here, $M = 5$ and $L = 3$

Consider a **total of N trials** in which we sample both variable X and Y

- Let n_{ij} be the number of such trials in which $X = x_i$ and $Y = y_j$
- Let c_i be the number of trials in which X takes the value x_i (irrespective of the value that Y takes)
- Let r_j be the number of trials in which Y takes the value y_j (irrespective of the value that X takes)

Probability theory (cont.)

The probability that X will take the value x_i and Y will take the value y_j is written $p(X = x_i, Y = y_j)$: It is the **joint probability** of $X = x_i$ and $Y = y_j$



It is given by the number of points falling in the cell (i, j) as a fraction of the total number N of points

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} \qquad (1)$$

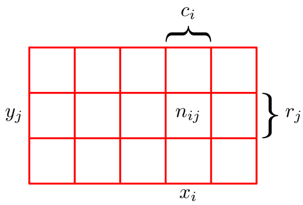
We are implicitly in the limit $N \rightarrow \infty$

Probability theory (cont.)

The probability that X takes the value x_i irrespective of the value of Y is $p(X = x_i)$ and is given by the fraction of total number of points in column i

$$p(X = x_i) = \frac{c_i}{N} = \frac{\sum_{j=1}^{L} n_{ij}}{N} = \sum_{j=1}^{L} \underbrace{\frac{n_{ij}}{N}}_{p(X=x_i, Y=y_j)} = \sum_{j=1}^{L} p(X = x_i, Y = y_j) \tag{2}$$

$p(X = x_i)$ is called the **marginal probability** because it obtained by marginalising, or summing out, the other variables (i.e., Y)



Definition

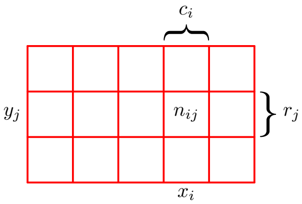
The definition of marginal probability sets the **Sum rule** of probability

$$p(X = x_i) = \sum_{j=1}^{L} p(X = x_i, Y = y_j) \tag{3}$$

Probability theory (cont.)

Consider only those instances for which $X = x_i$

- the fraction of such instances for which $Y = y_j$ is $p(Y = y_j | X = x_i)$
- It is the **conditional probability** of $Y = y_j$, given $X = x_i$



It is obtained by finding the fraction of points in column i that fall in cell i, j

$$p(Y = y_i | X = x_i) = \frac{n_{ij}}{c_i} \qquad (4)$$

Definition

From Equation 1, 2 and 4, we derive the **Product rule** of probability

$$\begin{aligned} p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} &= \underbrace{\frac{n_{ij}}{c_i}}_{p(Y=y_j|X=x_i)} \underbrace{\frac{c_i}{N}}_{p(X=x_i)} \\ &= p(Y = y_j | X = x_i) p(X = x_i) \end{aligned} \qquad (5)$$

Probability theory (cont.)

Definition

The rules of probability

- **sum rule**

$$p(X) = \sum_{Y} p(X, Y) \tag{6}$$

- **product rule**

$$p(X, Y) = p(Y|X)p(X) \tag{7}$$

To compact notation, $p(\star)$ denotes a distribution over a random variable $\star$ [1]

- $p(X, Y)$ is a joint probability, the probability of X and Y
- $p(Y|X)$ is a conditional probability, the probability of Y given X
- $p(X)$ is a marginal probability, the probability of X

[1] $p(\star = \cdot)$ or simply $p(\cdot)$ denotes the distribution evaluated for the particular value $\cdot$

Probability theory (cont.)

Definition

From the product rule and the symmetry property $p(X, Y) = p(Y, X)$, we obtain the following relationship between conditional probabilities

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \tag{8}$$

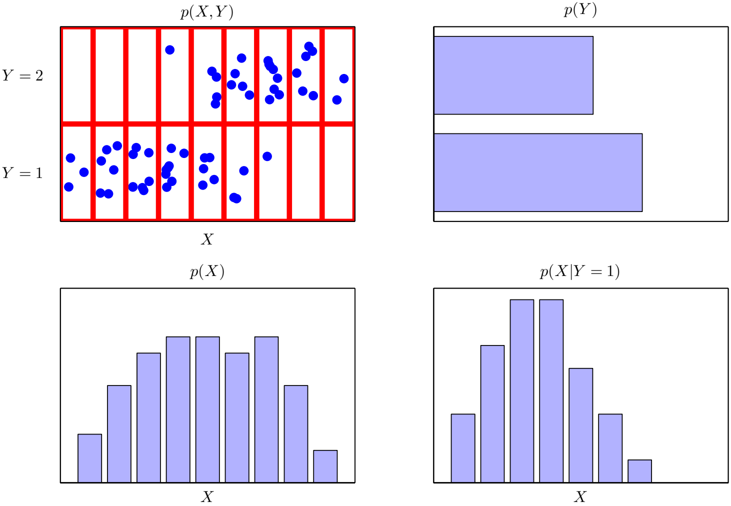
It is the **Bayes' theorem**, plays a central role in statistical machine learning

Using the sum rule, the denominator in Bayes' theorem can be expressed in terms of the quantities appearing in the numerator

$$p(X) = \sum_{Y} p(X|Y)p(Y) \tag{9}$$

The denominator is a normalisation constant that ensures that the sum of the conditional probability on the left-hand side of Eq. 8 over all values of Y is one

Probability theory (cont.)



Probability theory (cont.)

Example

Returning to the example involving the boxes of fruit

The probability of selecting either red or blue boxes are

- $p(B = r) = 4/10$ and $p(B = b) = 6/10$

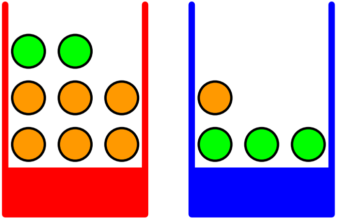
This satisfies $p(B = r) + p(B = b) = 4/10 + 6/10 = 1$

Now suppose that we pick a box at random, say the blue box

Then the probability of selecting an apple is just the fraction of apples in the blue box which is 3/4, so $p(F = a|B = b) = 3/4$

Probability theory (cont.)

We can write all conditional probabilities for the type of fruit, given the box



$$
\begin{aligned}
p(F = a|B = r) &= 1/4 \quad (10)\\
p(F = o|B = r) &= 3/4 \quad (11)\\
p(F = a|B = b) &= 3/4 \quad (12)\\
p(F = o|B = b) &= 1/4 \quad (13)
\end{aligned}
$$

Note that these probabilities are normalised so that

$$p(F = a|B = r) + p(F = o|B = r) = 1 \qquad (14)$$

$$p(F = a|B = b) + p(F = o|B = b) = 1 \qquad (15)$$

Probability theory (cont.)

We can now use the sum and product rules of probability to evaluate the overall probability of choosing an apple [2]

$$\begin{aligned} p(F=a) &= p(F=a|B=r)p(B=r) + p(F=a|B=b)p(B=b) \\ &= \frac{1}{4}\times\frac{4}{10} + \frac{3}{4}\times\frac{6}{10} = \frac{11}{20} \end{aligned} \tag{16}$$

from which it follows (sum rule) that $p(F=o) = 1 - 11/20 = 9/20$

[2] $P(X) = \sum_Y p(X,Y)$ with $p(X,Y) = p(Y|X)p(X) = p(Y,X) = p(X|Y)p(Y)$

Probability theory (cont.)

Suppose instead we are told that a piece of fruit has been selected and it is an orange, and we would like to know which box it came from

We want the probability distribution over boxes conditioned on the identity of the fruit ($P(B|F)$), whereas the probabilities in Eq. 10-13 give the probability distribution over fruits conditioned on the identity of the box ($P(F|B)$)

We solve the problem of reversing the conditional probability (Bayes' theorem)

$$p(B = r|F = o) = \frac{p(F = o|B = r)p(B = r)}{p(F = o)} = \frac{3}{4} \times \frac{4}{10} \times \frac{20}{9} = \frac{2}{3} \qquad (17)$$

From which it follows (sum rule) that $p(B = b|F = o) = 1 - 2/3 = 1/3$

Probability theory (cont.)

We can provide an important interpretation of Bayes' theorem as follows

- If we had been asked which box had been chosen before being told the identity of the selected item of fruit, then the most complete information we have available is provided by the probability $p(B)$
- We call this the **prior probability** because it is the probability available before we observe the identity of the fruit
- Once we are told that the fruit is an orange, we can then use Bayes' theorem to compute the probability $p(B|F)$
- We call this the **posterior probability** because it is the probability obtained after we have observed the identity of the fruit

The prior probability of selecting the red box was $4/10$ (the blue box is more probable), and once we observed that the piece of selected fruit is an orange, the posterior probability of the red box is $2/3$ (the red box is more probable)

Probability theory (cont.)

If the joint distribution of two variables factorises into the product of the marginals, $p(X, Y) = p(X)p(Y)$, then X and Y are said to be **independent**

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

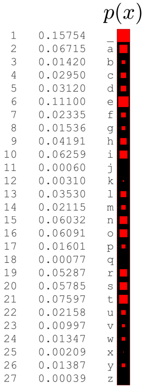
From the product rule, we see that $p(Y|X) = p(Y)$, and so the conditional distribution of Y given X is indeed independent of the value of X

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

Probability theory (cont.)

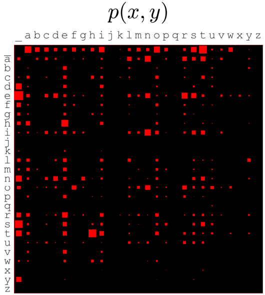
Example

An English language book (C. R. Darwin: *On the origin of species*, 1859)



$p(x)$

1	0.15754	_
2	0.06715	a
3	0.01420	b
4	0.02950	c
5	0.03120	d
6	0.11100	e
7	0.02335	f
8	0.01536	g
9	0.04191	h
10	0.06259	i
11	0.00060	j
12	0.00310	k
13	0.03530	l
14	0.02115	m
15	0.06032	n
16	0.06091	o
17	0.01601	p
18	0.00077	q
19	0.05287	r
20	0.05785	s
21	0.07597	t
22	0.02158	u
23	0.00997	v
24	0.01347	w
25	0.00209	x
26	0.01387	y
27	0.00039	z

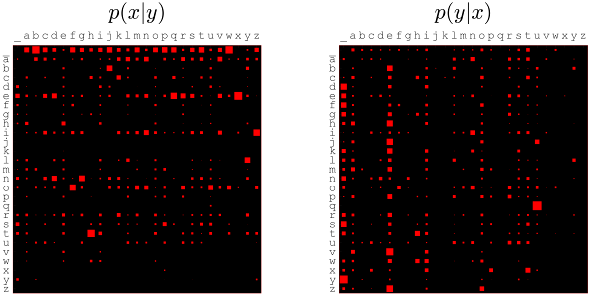


$p(x, y)$

- The probability distribution over the 27 possible letters
- The probability distribution over the 27 × 27 bigrams

Probability theory (cont.)

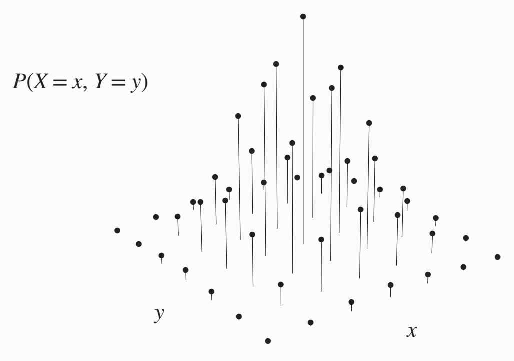
Example



- The conditional probability distribution of the first letter, given the second letter in a bigram
- The conditional probability distribution of the second letter, given the first one in a bigram *xy*

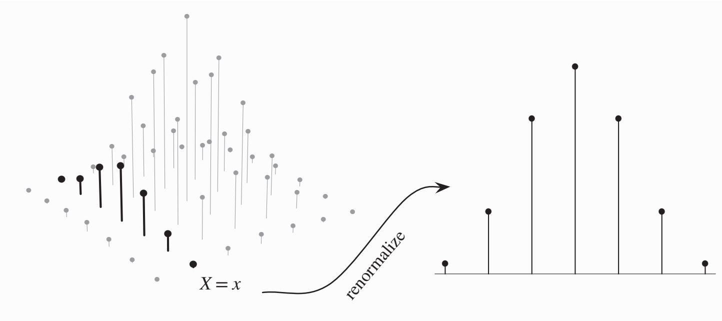
Probability theory (cont.)

Example



Probability theory (cont.)

Example



Probability densities

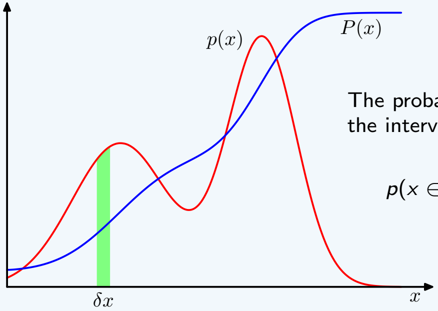
Probability theory

Probability densities

We wish now to consider probabilities with respect to continuous variables

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

Definition



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

$$p(x \in (a, b)) = \int_a^b p(x) dx \qquad (18)$$

Probability densities (cont.)

A more intuitive interpretation of the density function may be obtained from

$$p(x \in (a - \frac{\delta x}{2}, a + \frac{\delta x}{2})) = \int_{a-\delta x/2}^{a+\delta x/2} p(x) dx \approx \delta x p(a)$$

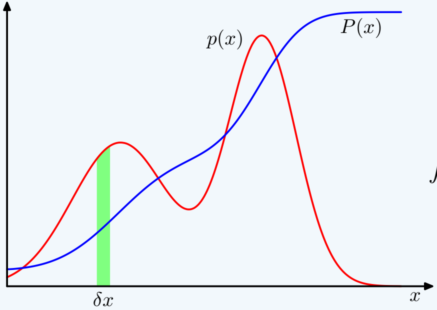
The probability that x is in a δx-wide interval around a is approximately $\delta x p(a)$

- $p(a)$ is a measure of how likely it is that random variable x will be near a

Probability densities (cont.)

Probabilities are nonnegative, and because the value of x must lie somewhere on the real axis, the probability density $p(x)$ must satisfy the two conditions

Definition



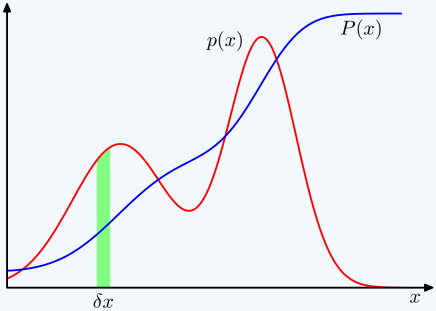
$$p(x) \geq 0 \qquad (19)$$

$$\int_{-\infty}^{+\infty} p(x)dx = 1 \qquad (20)$$

Probability densities (cont.)

The probability that x lies in the interval $(-\infty, z)$ is given by the **cumulative distribution function**, which is defined by

Definition



$$P(x) = \int_{-\infty}^{a} p(x)dx \qquad (21)$$

Density $p(x)$ is the derivative of the cumulative distribution function $P(x)$:

$$P'(x) = p(x), \quad \text{or } \frac{d}{dx}P(x) = p(x)$$

Probability densities (cont.)

Consider several continuous variables $x_1, \ldots, x_D$, collected in vector $\mathbf{x}$

- We define a **joint probability density** $p(\mathbf{x}) = p(x_1, \ldots, x_D)$ such that the probability of $\mathbf{x}$ falling in an infinitesimal volume $\delta\mathbf{x}$ containing $\mathbf{x}$ is $p(\mathbf{x})\delta\mathbf{x}$

Remark

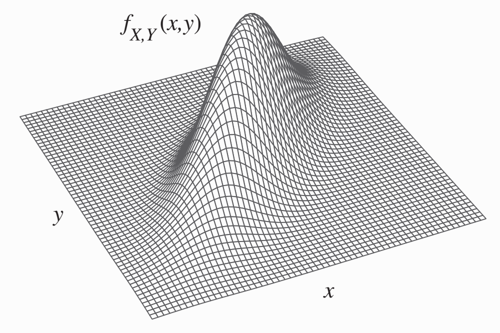
Also the multivariate probability density must satisfy

$$p(\mathbf{x}) \geq 0 \tag{22}$$

$$\int p(\mathbf{x}) d\mathbf{x} = 1 \tag{23}$$

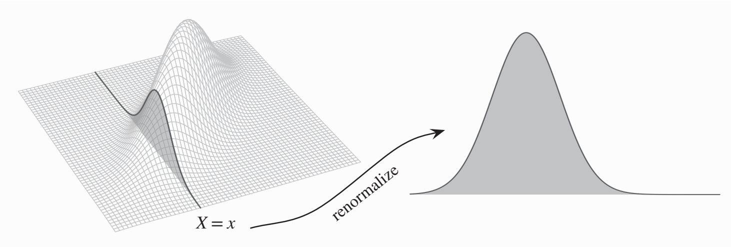
Probability densities (cont.)

Example



Probability densities (cont.)

Example



Probability densities (cont.)

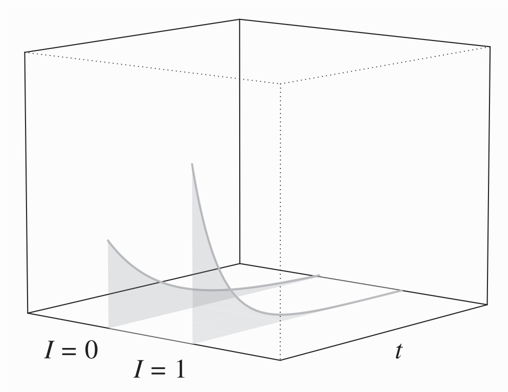
We can also consider joint probability distributions over a combination of discrete and continuous variables

Remark

Note that if x is a discrete variable, then $p(x)$ is sometimes called a **probability mass function** because it can be regarded as a set of ‘probability masses’ concentrated at the allowed values of x

Probability densities (cont.)

Example



Probability densities (cont.)

The sum and product rules, and Bayes' theorem, apply to the case of probability densities, or to combinations of discrete/continuous variables

Remark

If x and y are two real variables, the sum and product rules take the form

$$p(x) = \int p(x, y) dy \qquad (24)$$

$$p(x, y) = p(y|x)p(x) \qquad (25)$$

Expectations and covariances

Probability theory

Expectations and covariances

One operation involving probabilities is finding weighted averages of functions

- 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 $\mathbb{E}[f]$

Definition

For a discrete distribution, it is given by

$$\mathbb{E}[f] = \sum_x p(x) f(x) \tag{26}$$

so that the average is weighted by the relative probabilities of the values of x

Definition

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

$$\mathbb{E}[f] = \int p(x) f(x) dx \tag{27}$$

Expectations and covariances (cont.)

In either case, if we are given a finite number N of points drawn from the probability distribution or probability density, then the expectation can be approximated as a finite sum over these points

$$\mathbb{E}[f] \simeq \frac{1}{N}\sum_{n=1}^{N} f(x_n) \tag{28}$$

The approximation becomes exact in the limit $N \rightarrow \infty$

Expectations and covariances (cont.)

Sometimes we will be considering expectations of functions of several variables

- we can use a subscript to indicate which variable is being averaged over

Definition

$\mathbb{E}_x[f(x, y)]$ denotes the average of function $f(x, y)$ wrt the distribution of x

- $\mathbb{E}_x[f(x, y)] = \sum_x p(x) f(x, y)$
- $\mathbb{E}_x[f(x, y)]$ is a function of y

Expectations and covariances (cont.)

Definition

We can also consider a **conditional expectation** wrt a conditional distribution

$$\mathbb{E}_x[f(x)|y] = \sum_x p(x|y)f(x) \tag{29}$$

with an analogous definition for continuous variables

Expectations and covariances (cont.)

Definition

The measure of how much variability there is in $f(x)$ around its mean $\mathbb{E}[f(x)]$ is called the **variance** of $f(x)$ and it is defined by

$$\text{var}[f] = \mathbb{E}\Big[\Big(f(x) - \mathbb{E}[f(x)]\Big)^2\Big] \tag{30}$$

Expanding the square, we can show $(\star)$ that the variance can also be written in terms of the expectations of $f(x)$ and $f(x)^2$

$$\text{var}[f] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2 \tag{31}$$

- The variance of the variable x itself (i.e., $f(x) = x$) is

$$\text{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 \tag{32}$$

Expectations and covariances (cont.)

Definition

For two random variables x and y, the extent to which x and y vary together

- It is called **covariance** and it is defined by

$$\begin{aligned} \text{cov}[x, y] &= \mathbb{E}_{xy}\Big[(x - \mathbb{E}[x])(y - \mathbb{E}[y])\Big] \\ &= \mathbb{E}_{xy}[xy] - \mathbb{E}[x]\mathbb{E}[y] \end{aligned} \tag{33}$$

If x and y are independent, then their covariance vanishes ($\star$)

For two vectors of random variables $\mathbf{x}$ and $\mathbf{y}$, the covariance is a matrix

$$\begin{aligned} \text{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\Big[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{y}^T - \mathbb{E}[\mathbf{y}^T])\Big] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^T] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^T] \end{aligned} \tag{34}$$

Bayesian probabilities
Probability theory

Bayesian probabilities

We viewed probabilities as frequencies of repeatable random events

- It is the **frequentist** interpretation of probability

In general, we can view probabilities as quantification of uncertainty

- It is the **Bayesian** interpretation of probability

Example

In the example of the boxes of fruit the observation of the identity of the fruit provided relevant information that altered the probability of the chosen box

- Bayes's theorem converted a prior probability ($P(B = r) = 4/10$) into a posterior probability by incorporating the evidence by the observed data

$$p(B = r|F = o) = \frac{p(F = o|B = r)p(B = r)}{p(F = o)} = \frac{2}{3}$$

Bayesian probabilities (cont.)

We can adopt a similar approach when making inference about any quantities

- The parameters $\mathbf{w}$ in the polynomial curve fitting example

Definition

- We first capture our assumptions about $\mathbf{w}$, before observing the data in the form of a prior probability $p(\mathbf{w})$
- The effect of the observed data $\mathcal{D} = \{t_1, \dots, t_n\}$ is expressed through the conditional probability $p(\mathcal{D}|\mathbf{w})$
- We evaluate the uncertainty in $\mathbf{w}$, after we have observed $\mathcal{D}$ in the form of the posterior probability $p(\mathbf{w}|\mathcal{D})$

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})} \tag{35}$$

Bayesian probabilities (cont.)

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$$

The quantity $p(\mathcal{D}|\mathbf{w})$ is evaluated for the observed $\mathcal{D}$ and can be viewed as a function of the parameter vector $\mathbf{w}$, as such it is known as **likelihood function**

- It expresses how probable $\mathcal{D}$ is for different settings of the parameters $\mathbf{w}$

Bayesian probabilities (cont.)

The likelihood function $p(\mathcal{D}|\mathbf{w})$ plays a fundamental role

- In a frequentist setting, $\mathbf{w}$ is considered as a fixed parameter, whose value is determined by some form of *estimator*, and error bars on this estimate are obtained by considering the distribution of possible data sets $\mathcal{D}$
- In the Bayesian setting, there is only a single data set $\mathcal{D}$ (namely the one that is actually observed), and the uncertainty in the parameters is expressed through a probability distribution over $\mathbf{w}$ given that data set

Remark

The likelihood $p(\mathcal{D}|\mathbf{w})$ is NOT a probability distribution

Bayesian probabilities (cont.)

A widely used frequentist estimator is **maximum likelihood**, in which $\mathbf{w}$ is set to the value that maximises the likelihood function $p(\mathcal{D}|\mathbf{w})$

- This corresponds to choosing the value of $\mathbf{w}$ for which the probability of the observed data set $\mathcal{D}$ is maximised

Definition

The negative log of the likelihood function is called an **error function**

- The negative logarithm is a monotonically decreasing function, maximising the likelihood is equivalent to minimising the error

Bayesian probabilities (cont.)

Definition

Given the definition of likelihood, we state the Bayes' theorem also in words

posterior $\propto$ **likelihood** $\times$ **prior**

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})} \tag{36}$$

where all quantities are intended as functions of **w** and the denominator is a normalisation constant ensuring that the posterior distribution is a valid pdf

Integrating both sides of the Bayes' theorem with respect to **w**, we can express the denominator in terms of the prior distribution and the likelihood function

$$p(\mathcal{D}) = \int p(\mathcal{D}|\mathbf{w})p(\mathbf{w})d\mathbf{w} \tag{37}$$

The Gaussian distribution

Probability theory

The Gaussian distribution

We introduce an important probability distribution for continuous variables

- The **normal** or **Gaussian distribution**

Definition

For a single real-valued variable x, the Gaussian distribution is defined by

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\Big(-\frac{1}{2\sigma^2}(x-\mu)^2\Big) \tag{38}$$

It is a function of the variable x and it is governed by two parameters

- μ, called the **mean**
- σ^2 called the **variance**

The square root σ of the variance is the **standard deviation**

The reciprocal $\beta = \dfrac{1}{\sigma^2}$ of the variance is called the **precision**

The Gaussian distribution (cont.)

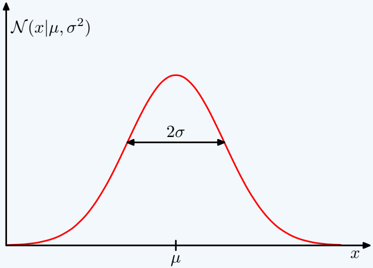
From the plot, the univariate Gaussian with mean μ and standard deviation σ

$$\mathcal{N}(x|\mu,\sigma) > 0 \tag{39}$$

In addition, the Gaussian distribution is normalised ($\star$)

$$\int_{-\infty}^{+\infty} \mathcal{N}(x|\mu,\sigma) = 1 \tag{40}$$

Remark



The Gaussian satisfies the two requirements for a valid probability density

The Gaussian distribution (cont.)

We can find expectations of functions of x under the Gaussian ($\star$)

- The average value of x is

$$\mathbb{E}[x] = \int_{-\infty}^{+\infty} \mathcal{N}(x|\mu,\sigma) x dx = \mu \tag{41}$$

- The second order moment

$$\mathbb{E}[x^2] = \int_{-\infty}^{+\infty} \mathcal{N}(x|\mu,\sigma) x^2 dx = \mu^2 + \sigma^2 \tag{42}$$

From Equation 41 and 42 follows that the variance of x is

$$\text{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2 \tag{43}$$

The Gaussian distribution (cont.)

Definition

The maximum of a distribution is called **mode** and for the Gaussian it is found in the correspondence of the mean ($\star$)

Definition

An important fact about normal random variables

- If variable x is normally distributed with parameters μ and σ^2, then $y = \alpha x + \beta$ is normally distributed with parameters $\alpha\mu + \beta$ and $\alpha^2\sigma^2$

The Gaussian distribution (cont.)

Definition

The Gaussian defined over a D-dimensional vector $\mathbf{x}$ of continuous variables

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\sigma}^2) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\Big(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big) \qquad (44)$$

- the D-dimensional vector $\boldsymbol{\mu}$ is the mean
- the $D \times D$ matrix $\boldsymbol{\Sigma}$ is the covariance
- $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$

The Gaussian distribution (cont.)

Example

We have a dataset $\mathtt{x} = (x_1, \dots, x_N)^T$ of N observations of a scalar variable x

- The observations are drawn independently from a Gaussian distribution
- The mean μ and variance σ^2 of the Gaussian distribution are unknown

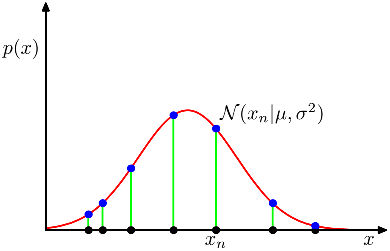
We know that the joint probability of two independent events equals the product of the marginal probabilities for each event separately

- Our data $\mathtt{x}$ are independently drawn from the same distribution (iid)
- We can write the probability of the data as a whole, given μ and σ^2

$$p(\mathtt{x}|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2) \tag{45}$$

The Gaussian distribution (cont.)

Seen as a function of μ and σ^2, this is the **likelihood function** for the Gaussian



The Gaussian distribution (red curve)

The black points denote a data set of values $\{x_n\}$

The likelihood function is the product of the blue values

Remark

One criterion for finding the parameters in a probability distribution using an observed set of data is to find the parameters that **maximise the likelihood**

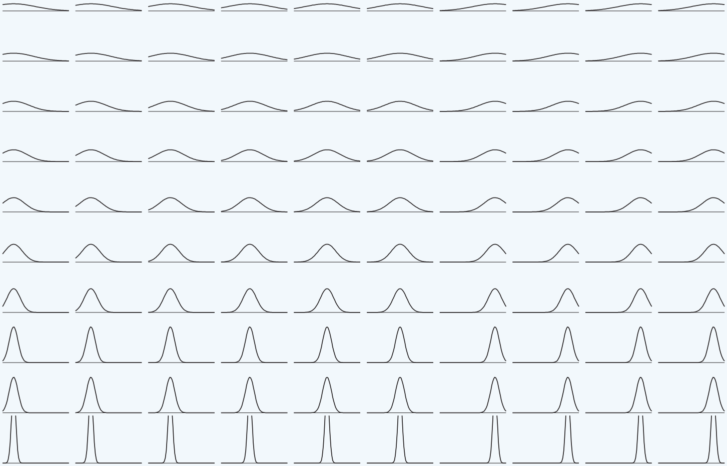
- Here, maximising the likelihood involves adjusting mean and variance

The Gaussian distribution (cont.)

Example

An entire (discretised) hypothesis space for a Gaussian, parameters μ and σ^2

- μ, horizontal axis
- σ, vertical axis

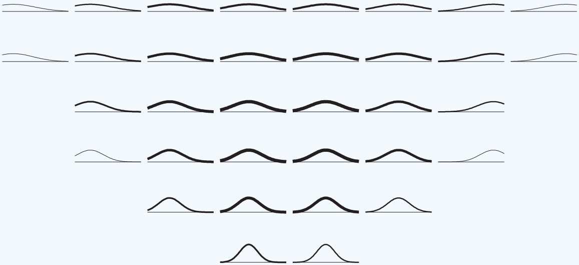


The Gaussian distribution (cont.)

Example

The likelihood function, given the data, as line thickness on the Gaussian

- Sub-hypothesis with likelihood larger than $1e^{-8}$ of max likelihood



The Gaussian distribution (cont.)

Instead of finding the values for the parameters μ and σ^2 in the Gaussian by maximising the likelihood, it is more convenient to maximise its logarithm[3]

- It simplifies the subsequent mathematics and helps numerically[4]

From Equation 38 and 45, the log likelihood can be written as

$$\ln p(\mathtt{x}|\mu, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2 - \frac{N}{2}\ln \sigma^2 - \frac{N}{2}\ln(2\pi) \qquad (46)$$

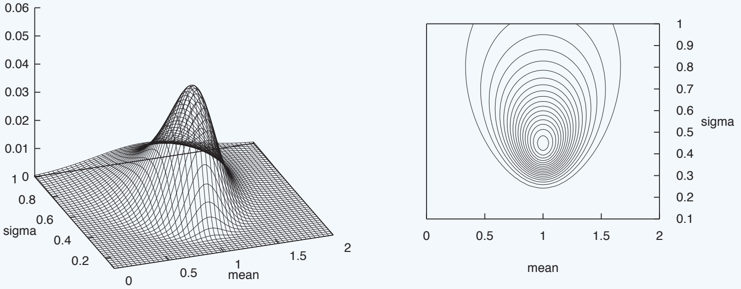
[3] Because the logarithm is a monotonically increasing function of its argument, maximisation of the log of a function is equivalent to maximisation of the function itself

[4] The product of a large number of small probabilities can easily overflow the numerical precision of the computer and this is resolved by calculating sums of the log probabilities

The Gaussian distribution (cont.)

Example

The likelihood function for the parameters of a Gaussian distribution



Surface plot and contour plot of log-likelihood, as function of μ and σ

The Gaussian distribution (cont.)

Definition

Maximising wrt μ, we get the maximum likelihood solution for the mean ($\star$)

- The **sample mean**

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{47}$$

Definition

Maximising wrt σ^2, we get the maximum likelihood solution for the variance

- The **sample variance**

$$\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N} (x_n - \mu_{ML})^2 \tag{48}$$

Note that we have to perform the joint maximisation of the log likelihood (wrt both μ and σ^2) but in the case of the Gaussian the solution of μ decouples from that of σ^2 and we can first evaluate Eq. 47 and use the result in Eq. 48

The Gaussian distribution (cont.)

One of the limitations of solutions using the maximum likelihood setting is that the approach systematically underestimates the variance of the distribution

- It is an example of a phenomenon called **bias** (relates to over-fitting)

Maximum likelihood solutions μ_{ML} and σ^2_{ML} are functions of $x_1, \ldots, x_n$

Definition

If we consider the expectations of these quantities wrt to the data (also from a Gaussian with parameters μ and σ^2) we can show ($\star$) that

$$\mathbb{E}[\mu_{ML}] = \mu \tag{49}$$

$$\mathbb{E}[\sigma^2_{ML}] = \Big(\frac{N-1}{N}\Big)\sigma^2 \tag{50}$$

so that on average the maximum likelihood estimate will obtain the correct mean but will underestimate the true variance by a factor $(N - 1/N)$

The Gaussian distribution (cont.)

Definition

From Eq. 50 it follows that an unbiased estimate of the variance parameter is

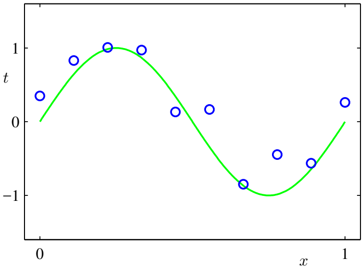
$$\tilde{\sigma}^2 = \frac{N}{N-1}\sigma_{ML}^2 = \frac{1}{N-1}\sum_{n-1}^{N}(x_n - \mu_{ML})^2 \tag{51}$$

Note that the bias of the maximum likelihood solution would anyway become less significant as the number of points N increases, and for $N \to \infty$ the solution equals the true variance of the distribution that generated the data

Polynomial fitting

Probability theory

Polynomial fitting



Training points $N = 10$ (blue circles)

- each comprising an observation of the **input variable** x along with corresponding **target variable** t

The **unknown function** $\sin(2\pi x)$ is used to generate data (green curve)

- Goal: Predict the value of t for some new value of x
- w/o knowledge of green curve

The **input training data** $\mathtt{x}$ was generated by choosing values of x_n, for $n = 1, \dots, N$, that are spaced uniformly in the range $[0, 1]$

The **target training data** $\mathtt{t}$ was obtained by computing values $\sin(2\pi x_n)$ of the function and adding a small level of Gaussian noise

Polynomial fitting (cont.)

- We shall fit the data using a polynomial function of the form

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j \qquad (52)$$

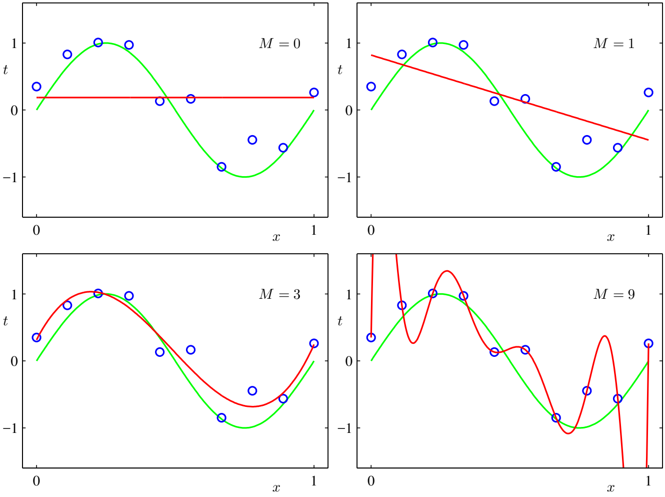
- M is the polynomial order, x^j is x raised to the power of j
- Polynomial coefficients $w_0, \dots, w_M$ are collected in vector $\mathbf{w}$

The coefficients values are obtained by fitting the polynomial to training data

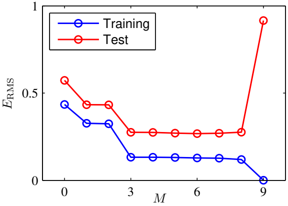
- By minimising an **error function**, a measure of misfit between function $y(x, \mathbf{w})$, for any given value of $\mathbf{w}$, and the training set data points
- A choice of error function is the sum of the squares of the errors between predictions $y(x_n, \mathbf{w})$ for each point x_n and corresponding target values t_n

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \Big(y(x_n, \mathbf{w}) - t_n \Big)^2 \quad \Longrightarrow \quad \mathbf{w}^{\star} \qquad (53)$$

Polynomial fitting (cont.)



Polynomial fitting (cont.)



Definition

The root mean squared error E_{RMS}

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

The magnitude of the coeffs tends to explode trying to (over)fit the data

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

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

Polynomial fitting (cont.)

One technique that is often used to control over-fitting is **regularisation**

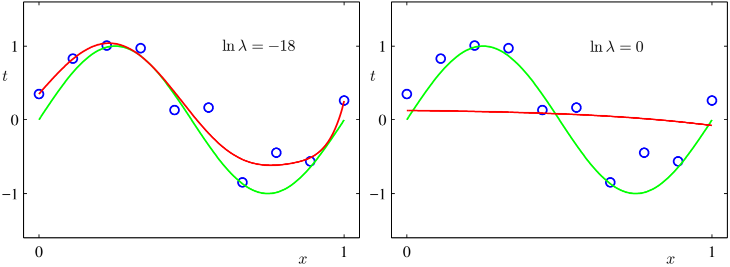
- Add a penalty term to the error function $E(\mathbf{w})$, to discourage the coefficients from reaching large values
- The simplest such penalty term is the sum of squares of all of the coefficients, to get a new error function

$$\tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\Big(y(x_n, \mathbf{w}) - t_n\Big)^2 + \frac{\lambda}{2}||\mathbf{w}||^2 \qquad (54)$$

- where $||\mathbf{w}||^2 = \mathbf{w}^T\mathbf{w} = w_0^2 + w_1^2 + \cdots + w_M^2$
- Coefficient λ trades off t between the regularisation term and the standard sum-of-squares error

Polynomial curve fitting (cont.)

Fitting the polynomial of order $M = 9$ to the data using a regularised error



- For $\ln \lambda = -18$ (it's a small value for λ), over-fitting is suppressed
- For $\ln \lambda = 0$ (it's a large value for λ), we obtain again a poor fit

Polynomial fitting (cont.)

Example

We have expressed the problem of polynomial curve fitting

$$\text{Error minimisation} \Rightarrow \begin{cases} E(\mathbf{w}) = \dfrac{1}{2}\sum_{n=1}^{N} \Big(y(x_n, \mathbf{w}) - t_n\Big)^2 \\ \\ \tilde{E}(\mathbf{w}) = \dfrac{1}{2}\sum_{n=1}^{N} \Big(y(x_n, \mathbf{w}) - t_n\Big)^2 + \dfrac{\lambda}{2}||\mathbf{w}||^2 \end{cases} \tag{55}$$

We return to it and view it from a probabilistic perspective

Polynomial fitting revisited

Polynomial fitting

Polynomial fitting revisited (cont.)

The goal in the curve fitting problem is to be able to make predictions for the target variable t, given some new value of the input variable x and

- a set of training data comprising N input values $\mathtt{x} = (x_1, \dots, x_N)^T$ and their corresponding target values $\mathtt{t} = (t_1, \dots, t_N)^T$

Uncertainty over the target value is expressed using a probability distribution

Assumption

- Given the value of x, the corresponding value of t is assumed to have a Gaussian distribution with a mean the value $y(x, \mathbf{w})$ of the polynomial

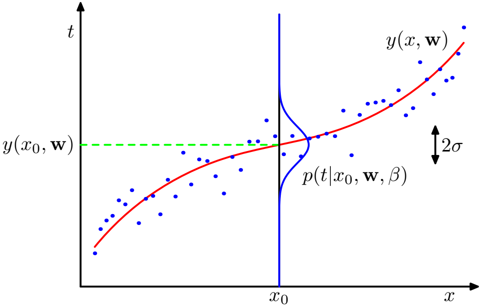
$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}\Big(t \Big| y(x, \mathbf{w}), \beta^{-1}\Big) \tag{56}$$

and some precision β (the precision is the reciprocal of the variance σ^2)

Polynomial fitting revisited (cont.)

The conditional distribution on t given x is $p(t|x, \mathbf{w}, \beta) = \mathcal{N}\Big(t\Big|y(x, \mathbf{w}), \beta^{-1}\Big)$

- The mean is given by the polynomial function $y(x, \mathbf{w})$
- The precision is given by β, with $\beta^{-1} = \sigma^2$



We can use training data $\{\mathtt{x}, \mathtt{t}\}$ to determine the values of the parameters μ and β of this Gaussian distribution

- **Likelihood maximisation**

Polynomial fitting revisited (cont.)

Assuming that the data have been drawn independently from the conditional distribution $p(t|x, \mathbf{w}, \beta) = \mathcal{N}\Big(t\Big|y(x, \mathbf{w}), \beta^{-1}\Big)$, the likelihood function is

$$p(\mathtt{t}|\mathtt{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\Big(t_n\Big|y(x_n, \mathbf{w}), \beta^{-1}\Big) \qquad (57)$$

It is again convenient to maximise its logarithm, the log likelihood function

$$\ln p(\mathtt{t}|\mathtt{x}, \mathbf{w}, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\Big(y(x_n, \mathbf{w}) - t_n\Big)^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) \qquad (58)$$

The optimisation is again with respect to both the polynomial coefficients $\mathbf{w}$ and the precision parameter β of the Gaussian conditional distribution

Polynomial fitting revisited (cont.)

$$\ln p(\mathtt{t}|\mathtt{x}, \mathbf{w}, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\Big(y(x_n, \mathbf{w}) - t_n\Big)^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)$$

Let us consider the determination of the maximum likelihood solution for **w**

- The last two terms can be omitted, as they do not depend on **w**
- Coefficient $\beta/2$ can be replaced with $1/2$, because scaling the log likelihood by a positive constant does not alter the location of its maximum with respect to **w**

Definition

Maximisation of log likelihood wrt **w** is minimisation of negative log likelihood

- This equals the minimisation of the sum-of-squares error function

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\Big(y(x_n, \mathbf{w}) - t_n\Big)^2 \quad \Longrightarrow \quad \mathbf{w}_{ML} = \mathbf{w}^{\star}$$

Polynomial fitting revisited (cont.)

$$\ln p(\mathtt{t}|\mathtt{x}, \mathbf{w}, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\Big(y(x_n, \mathbf{w}) - t_n\Big)^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)$$

Let us consider the determination of the maximum likelihood solution for β

Definition

- Maximising the log likelihood with respect to β gives

$$\frac{1}{\beta_{ML}} = \frac{1}{N}\sum_{n=1}^{N}\Big(y(x_n, \mathbf{w}_{ML}) - t_n)\Big)^2 \qquad (59)$$

- where again we decoupled the solution of $\mathbf{w}$ and β

Polynomial fitting revisited (cont.)

Having an estimate of **w** and β we can make predictions for new values of x

- We have a probabilistic model that gives the probability distribution over t

We can make estimations that are much more than a plain point estimate of t

- We can make predictions in terms of the **predictive distribution**

$$p(t|x, \mathbf{w}_{ML}, \beta_{ML}) = \mathcal{N}\Big(t\Big|y(x, \mathbf{w}_{ML}), \beta_{ML}^{-1}\Big) \tag{60}$$

- The probability distribution over t, rather than a point estimate

Polynomial fitting revisited (cont.)

We can make a step forward towards a Bayesian treatment of the problem

- We introduce a **prior distribution** over the polynomial coefficients $\mathbf{w}$

- We consider a Gaussian distribution[5]

$$\begin{aligned} \mathcal{N}(\mathbf{w}|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \mathcal{N}(\mathbf{w}|\mathbf{0},\alpha^{-1}\mathbf{I}) \\ &= \Big(\frac{\alpha}{2\pi}\Big)^{(M+1)/2}\exp\Big(-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\Big) = p(\mathbf{w}|\alpha) \end{aligned} \qquad (61)$$

- $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \alpha^{-1}\mathbf{I}$
- α is the precision of the distribution[6]
- Number of parameters in $\mathbf{w}$, $M+1$

[5] $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\sigma}^2) = \dfrac{1}{(2\pi)^{D/2}}\dfrac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\Big(-\dfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big)$

[6] Variables such as α control the distribution of model parameters are called **hyperparameters**

Polynomial fitting revisited (cont.)

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \Big(\frac{\alpha}{2\pi}\Big)^{(M+1)/2} \exp\Big(-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\Big)$$

$$p(\mathtt{t}|\mathtt{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\Big(t_n\Big|y(x_n, \mathbf{w}), \beta^{-1}\Big)$$

Using Bayes' theorem, the **posterior distribution** for **w** is proportional to the product of the prior distribution and the likelihood function, thus

$$p(\mathbf{w}|\mathtt{x}, \mathtt{t}, \alpha, \beta) \propto p(\mathtt{t}|\mathtt{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha) \qquad (62)$$

We can now determine **w** by finding its most probable value given the data

- that is, by **maximising the posterior distribution**
- this technique is **maximum posterior** or **MAP**

Polynomial fitting revisited (cont.)

By taking the negative log of the posterior distribution over **w** and combining with Eq. 58 (log likelihood function) and Eq. 61 (prior distribution over **w**), we find that the maximum of the posterior is given by the minimum of

$$\frac{\beta}{2}\sum_{n=1}^{N}\Big(y(x_n,\mathbf{w})-t_n\Big)^2+\frac{\alpha}{2}\mathbf{w}^T\mathbf{w} \tag{63}$$

Definition

Thus, maximising the posterior is equivalent to minimising the regularised sum-of-squares error function with regularisation $\lambda = \alpha/\beta$

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

Bayesian polynomial fitting

Polynomial fitting

Bayesian polynomial fitting (cont.)

Remark

Though we included a prior $p(\mathbf{w}|\alpha)$, we are still making point estimates of $\mathbf{w}$

- Not yet a full Bayesian treatment

In our problem, we are given training data $\mathtt{x}$ and $\mathtt{t}$, along with a new point x

- We wish to evaluate the **posterior predictive distribution** $p(t|x, \mathtt{x}, \mathtt{t})$

Assuming parameters α and β fixed and known, the predictive distribution is

$$p(t|x, \mathtt{x}, \mathtt{t}) = \int p(t|x, \mathbf{w}) p(\mathbf{w}|\mathtt{x}, \mathtt{t}) d\mathbf{w} \qquad (64)$$

- $p(t|x, \mathbf{w})$ is $p(t|x, \mathbf{w}, \beta) = \mathcal{N}\left(t \middle| y(x, \mathbf{w}), \beta^{-1}\right)$
- $p(\mathbf{w}|\mathtt{x}, \mathtt{t})$ is the posterior distribution over $\mathbf{w}$

$$p(\mathbf{w}|\mathtt{x}, \mathtt{t}, \alpha, \beta) \propto p(\mathtt{t}|\mathtt{x}, \mathbf{w}, \beta) p(\mathbf{w}|\alpha)$$

Bayesian polynomial fitting (cont.)

It is possible to show this posterior distribution is a Gaussian that can be evaluated analytically and also the integration can be performed analytically

$$p(t|x,\mathtt{x},\mathtt{t}) = \int p(t|x,\mathbf{w})p(\mathbf{w}|\mathtt{x},\mathtt{t})d\mathbf{w} = \mathcal{N}\Big(t\Big|m(x), s^2(x)\Big) \qquad (65)$$

The mean and variance of Gaussian posterior predictive distribution are

$$\begin{aligned} m(x) &= \beta\phi(x)^T\mathbf{S}\sum_{n=1}^{N}\phi(x_n)t_n && (66)\\ s^2(x) &= \beta^{-1}+\phi(x)^T\mathbf{S}\phi(x) && (67)\end{aligned}$$

We defined the vector $\phi(x)$ with elements $\phi_i(x) = x^i$, with $i = 0,\dots,M$

The matrix $\mathbf{S}$ is such that

$$\mathbf{S}^{-1} = \alpha\mathbf{I} + \beta\sum_{n=1}^{N}\phi(x_n)\phi(x_n)^T \qquad (68)$$

Bayesian polynomial fitting (cont.)

$$
\begin{aligned}
m(x) &= \beta \phi(x)^T \mathbf{S} \sum_{n=1}^{N} \phi(x_n) t_n \\
s^2(x) &= \beta^{-1} + \phi(x)^T \mathbf{S} \phi(x)
\end{aligned}
$$

We see that the variance, but also the mean, of this predictive distribution $p(t|x, \mathtt{x}, \mathtt{t}) = \mathcal{N}(t|m(x), s^2(x))$ depends on x

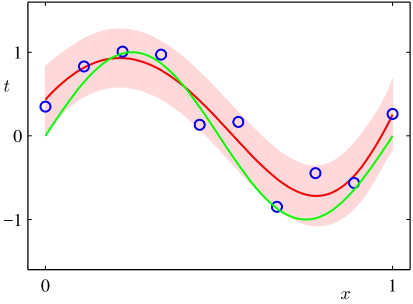
- The first terms in s^2 represents the uncertainty in the predicted value t due to the noise on the target variables
- It was already present in the maximum likelihood predictive distribution $p(t|x, \mathbf{w}_{ML}, \beta_{ML}) = \mathcal{N}(t|y(x, \mathbf{w}_{ML}), \beta_{ML}^{-1})$

The second term arises from the uncertainty in the parameters and it is a consequence of the Bayesian treatment

Bayesian polynomial fitting (cont.)

The predictive distribution $p(t|x, \mathtt{x}, \mathtt{t}) = \mathcal{N}\Big(t\Big|m(x), s^2(x)\Big)$, $M = 9$

- The red curve is the mean $m(x)$ of the predictive distribution
- The red region corresponds to ±1 s around the mean



- $\alpha = 5 \times 10^{-2}$
- $\beta = 11.1$, corresponding to the known noise variance

Graphical models
Probability theory

Graphical models

Probabilities play a central role in modern pattern recognition

Probability theory can be expressed in terms of two equations

- The sum rule and the product rule

$$\begin{aligned} p(X) &= \sum_Y p(X,Y) \\ p(X,Y) &= p(Y|X)p(X) \end{aligned}$$

All probabilistic inference and learning manipulations here, no matter how complex, amount to repeated application of these two equations

$$\begin{aligned} p(Y|X) &= \frac{p(X|Y)p(Y)}{p(X)} \\ p(X) &= \sum_Y p(X|Y)p(Y) \end{aligned}$$

We formulate and solve probabilistic models by algebraic manipulation

Graphical models (cont.)

It is advantageous to use visual displays of probability distributions

- **Probabilistic graphical models**

Probabilistic graphical models offer several useful properties:

- They provide a simple way to **visualise** the structure of a probabilistic model and can be used to design and motivate new models
- Insights into the properties of the model, including **conditional independence** properties, can be obtained by inspection of the graph
- The **computations** required to perform inference and learning in complex models, can be expressed in terms of graphical manipulations, in which underlying mathematical expressions are carried along implicitly

Graphical models (cont.)

A graph comprises **nodes** (or **vertices**) connected by **links** (or **edges** or **arcs**)

Definition

In a probabilistic graphical model

- each node represents a random variable (or group of random variables)
- the links express probabilistic relationships between these variables

The graph captures how the joint distribution over all random variables can be decomposed into a product of factors each depending only on variables' subset

Graphical models (cont.)

We discuss **Bayesian networks**, or **directed graphical models**, in which the links of the graphs have a particular directionality indicated by arrows

The other major class of graphical models are **Markov random fields**, or **undirected graphical models**, with links without directional significance

Bayesian networks
Graphical models

Bayesian networks

To motivate the use of directed graphs to describe probability distributions, consider any joint distribution $p(a, b, c)$ over three variables a, b, and c[7]

By using the product rule of probability, we can write the joint distribution as

$$p(a, b, c) = p(c|a, b)p(a, b) \tag{69}$$

A second application of the product rule to the second term on the RHS gives

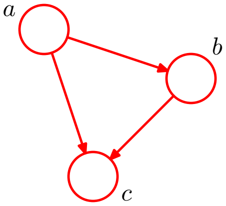
$$p(a, b, c) = p(c|a, b)\,\underbrace{p(b|a)p(a)}_{p(a,b)} \tag{70}$$

[7] We do not need to specify anything further about these variables (discrete, continuous, ...)

Bayesian networks (cont.)

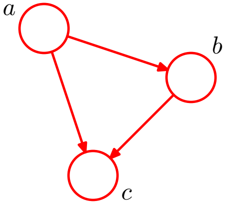
Decomposition $p(a, b, c) = p(c|a, b)p(b|a)p(a)$ holds for any joint distribution

We can represent the right-hand side as a probabilistic graphical model



1. We introduce a node for each of the random variables a, b, and c
2. We associate each node with the corresponding conditional distribution
3. For each conditional distribution, we add directed links from nodes corresponding to variables on which the distribution is conditioned

Bayesian networks (cont.)



$$p(a, b, c) = p(c|a, b)p(b|a)p(a)$$

- Factor $p(c|a, b)$, links from nodes a and b to node c
- Factor $p(b|a)$, links from node a to node b
- Factor $p(a)$, no incoming links

If there is a link from a node a to a node b, we say

- node a is the **parent** of node b
- node b is the **child** of node a

We make no formal distinction between node and variable it corresponds to

- We use the same symbol to refer to both

Bayesian networks (cont.)

The left-hand side of $p(a, b, c) = p(c|a, b)p(b|a)p(a)$ is symmetrical with respect to the three variables a, b, and c, whereas the right-hand side is not

Indeed, in making the decomposition, we chose a particular ordering (a, b, c)

- Had we chosen a different ordering, we would have obtained a different decomposition (and hence also a different graphical representation)
- Awkward

Bayesian networks (cont.)

Consider the joint distribution over K variables given by $p(x_1, \ldots, x_K)$

By repeated application of the product rule of probability, this joint distribution can be written as a product of conditional distributions, one for each variable

$$p(x_1, \ldots, x_K) = p(x_K|x_1, \ldots, x_{K-1})p(x_{K-1}|x_1, \ldots, x_{K-2}) \cdots p(x_2|x_1)p(x_1)$$

For a choice of K, we can represent this as a directed graph with K nodes

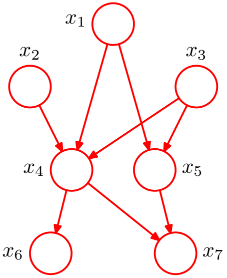
- one node for each conditional distribution on the right-hand side
- each node has incoming links from all lower numbered nodes

This graph is **fully connected**, there is a link between every pair of nodes

Bayesian networks (cont.)

We worked with general joint distributions, so that decompositions and their representations as fully connected graphs, are applicable to any distribution

It is the absence of links in the graph that conveys interesting information about the properties of the class of distributions that the graph represents



$$
\begin{aligned}
p(x_1, \cdots, x_7) \quad = \quad & p(x_7 | \cancel{x_1}, \cancel{x_2}, \cancel{x_3}, x_4, x_5, \cancel{x_6}) \\
& p(x_6 | \cancel{x_1}, \cancel{x_2}, \cancel{x_3}, x_4, \cancel{x_5}) \\
& p(x_5 | x_1, \cancel{x_2}, x_3, \cancel{x_4}) \\
& p(x_4 | x_1, x_2, x_3) \\
& p(x_3 | \cancel{x_1}, \cancel{x_2}) \\
& p(x_2 | \cancel{x_1}) \\
& p(x_1)
\end{aligned}
$$

$$
\begin{aligned}
& p(x_1, \ldots, x_7) = \\
& \qquad p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)
\end{aligned}
\tag{71}
$$

Bayesian networks (cont.)

We state in general terms the relationship between a given acyclic[8] directed graph and the corresponding distribution over the variables

- The joint distribution defined by a graph is given by the product, over all nodes of the graph, of a conditional distribution for each node conditioned on the variables corresponding to the parents of that node in the graph

Thus, for a graph with K nodes, the joint distribution is given by

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathbf{pa}_k) \tag{72}$$

where $\mathbf{pa}_k$ denotes the set of parents of x_k and $\mathbf{x} = \{x_1, \dots, x_K\}$

Factorisation properties of the joint distribution for a directed graphical model

[8] We consider directed graphs subjected to an important restriction: **No directed cycles**
In other words, there are no closed paths within the graph such that we can move from node to node along links following the direction of the arrows and end up back at the starting node

Bayesian polynomial fitting

Bayesian networks

Bayesian polynomial fitting

We illustrate the use of directed graphs to describe probability distributions

- Bayesian polynomial regression model

The random variables in this model are:

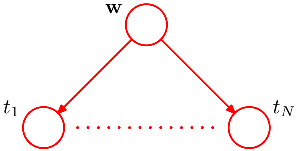
- the vector of polynomial coefficients **w**
- the observed data $=(t_1, \dots, t_N)^T$

The model contains input data $\mathtt{x} = (x_1, , \dots, x_N)^T$, the noise variance σ^2, and the hyper-parameter α representing the precision of the Gaussian prior over **w**

- All are parameters of the model, not random variables

Bayesian polynomial fitting (cont.)

The joint distribution of the random variables is given by the product of the prior $p(\mathbf{w})$ and N conditional distributions $p(t_n|\mathbf{w})$, with $n = 1, \dots, N$



$$p(\mathtt{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n|\mathbf{w}) \qquad (73)$$

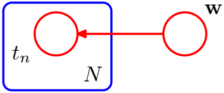
This joint distribution can represented by a probabilistic graphical model

It inconvenient to write out multiple nodes of the form $t_1, \dots, t_N$ explicitly

Bayesian polynomial fitting (cont.)

The graphical notation to compactly express multiple nodes is called **plate**

- We draw a single representative node t_n and then surround this with a box, labelled with N indicating that there are N nodes of this kind



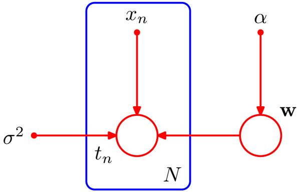
$$p(\mathtt{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n|\mathbf{w})$$

Bayesian polynomial fitting (cont.)

It is useful to make the parameters and its stochastic variables explicit

$$p(\mathtt{t}, \mathbf{w}|\mathtt{x}, \alpha, \sigma^2) = p(\mathbf{w}|\alpha) \prod_{n=1}^{N} p(t_n|\mathbf{w}, x_n, \sigma^2)$$

which allows to make $\mathtt{x}$ and α explicit in the graphical representation



The graphical convention

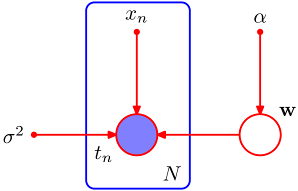
- Random variables are denoted by large open circles
- Deterministic parameters are denoted by small solid circles

Bayesian polynomial fitting (cont.)

Some of the random variables are set to the specific observed values

- In the example, variables $\{t_n\}$ from the training set

We denote **observed variables** by shading the corresponding nodes



Variables $\{t_n\}$ are observed, shaded

Variable $\mathbf{w}$ is not observed, unshaded

Variables that are not observed are called **hidden** or **latent variables**

Bayesian polynomial fitting (cont.)

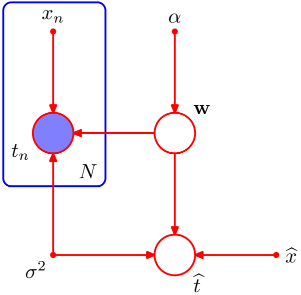
Because we observed the values $\{t_n\}$ we can evaluate the posterior distribution of the polynomial coefficients $\mathbf{w}$, it involves an application of Bayes' theorem

$$p(\mathbf{w}|\mathbf{T}) \propto p(\mathbf{w}) \prod_{n=1}^{N} p(t_n|\mathbf{w}) \qquad (74)$$

Bayesian polynomial fitting (cont.)

Model parameters like **w** are generally of little interest as such

Suppose we are given a new input $\hat{x}$ and we wish to find the corresponding probability distribution for target $\hat{t}$, conditioned on the observed data



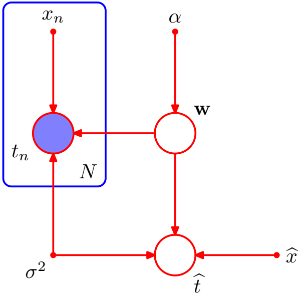
The joint distribution of all variables

- conditioned on parameters

$$p(\hat{t}, \mathtt{t}, \mathbf{w}|\hat{x}, \mathtt{x}, \alpha, \sigma^2) = \Big(\prod_{n=1}^{N} p(t_n|x_n, \mathbf{w}, \sigma^2)\Big) p(\mathbf{w}|\alpha) p(\hat{t}|\hat{x}, \mathbf{w}, \sigma^2) \qquad (75)$$

Bayesian polynomial fitting (cont.)

The joint distribution $p(\hat{t}, \mathtt{t}, \mathbf{w}|\hat{x}, \mathtt{x}, \alpha, \sigma^2)$ of the random variables, conditioned on the deterministic parameters is $\Big(\prod_{n=1}^{N} p(t_n|x_n, \mathbf{w}, \sigma^2)\Big) p(\mathbf{w}|\alpha) p(\hat{t}|\hat{x}, \mathbf{w}, \sigma^2)$



The (posterior) predictive distribution for $\hat{t}$ is obtained from the sum rule

- By integrating out the model parameters $\mathbf{w}$

$$p(\hat{t}|\hat{x}, \mathtt{x}, \mathtt{t}, \alpha, \sigma^2) \propto \int p(\hat{t}, \mathtt{t}, \mathbf{w}|\hat{x}, \mathtt{x}, \alpha, \sigma^2) d\mathbf{w} \tag{76}$$

The random variables in $\mathtt{t}$ are explicitly set to the specific observed values