Lecture notes for CSCI 6905: Autonomous Robotics September 2011

Thomas P. Trappenberg
Dalhousie University

1 Basic probability theory

A major milestone for the modern approach to machine learning is to acknowledge our limited knowledge about the world and the unreliability of sensors and actuators. It is then only natural to consider quantities in our approaches as **random variables**. While a regular variable, once set, has only one specific value, a random variable will have different values every time we 'look' at it (draw an example from the distributions). Just think about a light sensor. We might think that an ideal light sensor will give us only one reading while holding it to a specific surface, but since the peripheral light conditions change, the characteristics of the internal electronic might change due to changing temperatures, or since we move the sensor unintentionally away from the surface, it is more than likely that we get different readings over time. Therefore, even internal variables that have to be estimated from sensors, such as the state of the system, is fundamentally a random variable.

A common misconception about randomness is that one can not predict values of random values. Some values might be more likely than others, and, while we might not be able to predict a specific value when drawing a random number, it is possible to say something like how often a certain number will appear when drawing many examples. We might even be able to state how confident we are with this number, or, in other words, how variable these predictions are. The complete knowledge of a random variable, that is, how likely each value is for a random variable x, is captured by the **probability density function** $pdf(x)$. We discuss some specific examples of pdfs below. In these examples we assume that we know the pdf, but in may practical applications we must estimate this function. Indeed, estimation of pdfs is at the heart if not the central tasks of machine learning. If we would know the 'world pdf', the probability function of all possible events in the world, then we could predict as much as possible in this world.

Most of the systems discussed in this course are **stochastic models** to capture the uncertainties in the world. Stochastic models are models with random variables, and it is therefore useful to remind ourselves about the properties of such variables. This chapter is a refresher on concepts in probability theory. Note that we are mainly interested in the language of probability theory rather than statistics, which is more concerned with hypothesis testing and related procedures.

1.1 Random numbers and their probability (density) function

Probability theory is the theory of **random numbers**. We denoted such numbers by capital letters to distinguish them from regular numbers written in lower case. A random variable, X, is a quantity that can have different values each time the variable

is inspected, such as in measurements in experiments. This is fundamentally different to a regular variable, x, which does not change its value once it is assigned. A random number is thus a new mathematical concept, not included in the regular mathematics of numbers. A specific value of a random number is still meaningful as it might influence specific processes in a deterministic way. However, since a random number can change every time it is inspected, it is also useful to describe more general properties when drawing examples many times. The frequency with which numbers can occur is then the most useful quantity to take into account. This frequency is captured by the mathematical construct of a **probability**. Note that there is often a debate if random numbers should be defines solely on the basis of a frequency measurement, or if there they should be treated as a special kind of objects. This philosophical debate between 'Frequentists' and 'Bayesians' is of minor importance for our applications.

We can formalize the idea of expressing probabilities of drawing specific specific values for random variable with some compact notations. We speak of a **discrete random variable** in the case of discrete numbers for the possible values of a random number. A **continuous random variable** is a random variable that has possible values in a continuous set of numbers. There is, in principle, not much difference between these two kinds of random variables, except that the mathematical formulation has to be slightly different to be mathematically correct. For example, the **probability function**,

$$P_X(x) = P(X = x) \tag{1.1}$$

describes the frequency with which each possible value x of a discrete variable X occurs. Note that x is a regular variable, not a random variable. The value of $P_X(x)$ gives the fraction of the times we get a value x for the random variable X if we draw many examples of the random variable.[1] From this definition it follows that the frequency of having any of the possible values is equal to one, which is the normalization condition

$$\sum_x P_X(x) = 1. \tag{1.2}$$

In the case of continuous random numbers we have an infinite number of possible values x so that the fraction for each number becomes infinitesimally small. It is then appropriate to write the probability distribution function as $P_X(x) = p_X(x)\mathrm{d}x$, where $p_X(x)$ is the **probability density function** (pdf). The sum in eqn 1.2 then becomes an integral, and normalization condition for a continuous random variable is

$$\int_x p_X(x)\mathrm{d}x = 1. \tag{1.3}$$

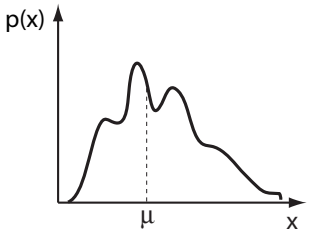
We will formulate the rest of this section in terms of continuous random variables. The corresponding formulas for discrete random variables can easily be deduced by replacing the integrals over the pdf with sums over the probability function. It is also possible to use the δ-function, outlined in Appendix **??**, to write discrete random processes in a continuous form.

[1] Probabilities are sometimes written as a percentage, but we will stick to the fractional notation.

1.2 Moments: mean, variance, etc.

In the following we only consider independent random values that are drawn from identical pdfs, often labeled as iid (independent and identically distributed) data. That is, we do not consider cases where there is a different probabilities of getting certain numbers when having a specific number in a previous trial. The static probability density function describes, then, all we can know about the corresponding random variable.

Let us consider the arbitrary pdf, $p_X(x)$, with the following graph:



Such a distribution is called **multimodal** because it has several peaks. Since this is a pdf, the area under this curve must be equal to one, as stated in eqn 1.3. It would be useful to have this function parameterized in an analytical format. Most pdfs have to be approximated from experiments, and a common method is then to fit a function to the data. We can also view this approximation as a learning problem, that is, how can we learn the pdf from data? We will return to this question later.

Finding a precise form of a pdf is difficult, and we became thus used to describing random variables with a small set of numbers that are meant to capture some properties. For example, we might ask what the most frequent value is when drawing many examples. This number is given by the largest peak value of the distribution. It is often more useful to know something about the average value itself when drawing many examples. A common quantity to know is thus the expected arithmetic average of those numbers, which is called the **mean**, **expected value**, or **expectation value** of the distribution, defined by

$$\mu = \int_{-\infty}^{\infty} xp(x)\mathrm{d}x. \tag{1.4}$$

This formula formalizes the calculation of adding all the different numbers together with their frequency.

A careful reader might have noticed a little oddity in our discussion. On the one hand we are saying that we want to characterize random variables through some simple measurements because we do not know the pdf, yet the last formula uses the pdf $p(x)$ that we usually don't know. To solve this apparent oddity we need to be more careful and talk about the **true underlying functions** and the **estimation** of such functions. If we would know the pdf that governs the random variable X, then equation 1.4 is the definition of the mean. However, in most applications we do not know the pdf, but we can define an approximation of the mean from measurements. For example, if we measure the frequency p_i of values in certain intervals around values x_i, then we can estimate the true mean μ by

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i p_i. \tag{1.5}$$

It is a common practice to denote an estimate of a quantity by adding a hat symbol to the quantity name. Also, not that we have use here a discretization procedure to approximate random variable that can be continuous in the most general case. Also note that we could enter here again the philosophical debate. Indeed, we have treated the pdf as fundamental and described the arithmetic average like an estimation of the mean. This might be viewed as *Bayesian*. However, we could also be pragmatic and say that we only have a collection of measurements so that the numbers are the 'real' thing, and that pdfs are only a mathematical construct. We will continue with a Bayesian description but note that this makes no difference at the end when using it in specific applications.

The mean of a distribution is not the only interesting quantity that characterizes a distribution. For example, we might want to ask what the **median** value is for which it is equally likely to find a value lower or larger than this value. Furthermore, the spread of the pdf around the mean is also very revealing as it gives us a sense of how spread the values are. This spread is often characterized by the standard deviation (std), or its square, which is called **variance**, σ^2, and is defined as

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\mathrm{d}x. \tag{1.6}$$

This quantity is generally not enough to characterize the probability function uniquely; this is only possible if we know all moments of a distribution, where the nth **moment about the mean** is defined as

$$m^n = \int_{-\infty}^{\infty} (x-\mu)^n f(x)\mathrm{d}x. \tag{1.7}$$

The **variance** is the second moment about the mean,

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\mathrm{d}x. \tag{1.8}$$

Higher moments specify further characteristics of distributions such as terms with third-order exponents (lie a quantity called skewness) or fourth-oder (such as a quantity called kurtosis). Moments higher than this have not been given explicit names. Knowing all moments of a distribution is equivalent to knowing the distribution precisely, and knowing a pdf is equivalent to knowing everything we could know about a random variable.

In case the distribution function is not given, moments have to be estimated from data. For example, the mean can be estimated from a sample of measurements by the **sample mean**,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{1.9}$$

and the variance from the **sample variance**,

$$s_1^2 = \frac{1}{n}\sum_{i=1}^{n} (\bar{x} - x_i)^2. \tag{1.10}$$

We will discuss later that these are the appropriate maximum likelihood estimates of these parameters. Note that the sample mean is an **unbiased estimate** while the sample variance is **biased**. A statistic is said to be biased if the mean of the sampling distribution is not equal to the parameter that is intended to be estimated. It can be shown that $E(s_1^2) = \frac{1}{n}\sigma^2$, and we can therefore adjust for the bias with a different normalization. It is hence common to use the **unbiased sample variance**

$$s_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{x} - x_i)^2, \tag{1.11}$$

as estimator of the variance.

Finally, it is good to realize that knowing all moments uniquely specifies a pdf. But the reverse is also true, that is, and incomplete list of moments does not uniquely define a pdf. Note that the all higher moments are zero for the Gaussian distributions, which means that the mean and variance uniquely define the distribution. This is however not the case for other distributions, and the usefulness of reporting these statistics can then be questioned.

1.3 Examples of probability (density) functions

There is an infinite number of possible pdfs. However, some specific forms have been very useful for describing some specific processes and have thus been given names. The following is a list of examples with some discrete and several continuous distributions. Most examples are discussed as one-dimensional distributions except the last example, which is a higher dimensional distribution.

1.4 Bernoulli distribution

A Bernoulli random variable is a variable from an experiment that has two possible outcomes: success with probability p; or failure, with probability $(1-p)$.

Probability function:
$P(\text{success}) = p; P(\text{failure}) = 1-p$
mean: p
variance: $p(1-p)$

1.4.1 Multinomial distribution

This is the distribution of outcomes in n trials that have k possible outcomes. The probability of each outcome is thereby p_i.

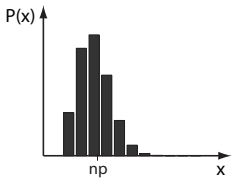
Probability function:
$P(x_i) = n! \prod_{i=1}^{k}(p_i^{x_i}/x_i!)$
mean: np_i
variance: $np_i(1-p_i)$

An important example is the Binomial distribution ($k = 2$), which describes the the

number of successes in n Bernoulli trials with probability of success p. Note that the binomial coefficient is defined as

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \tag{1.12}$$

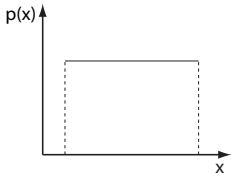
and is given by the MATLAB function `nchoosek`.



Probability function:
$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$
mean: np
variance: $np(1-p)$

1.4.2 Uniform distribution

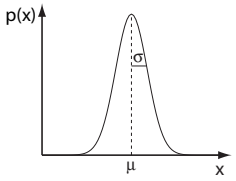
Equally distributed random numbers in the interval $a \leq x \leq b$. Pseudo-random variables with this distribution are often generated by routines in many programming languages.



Probability density function:
$p(x) = \frac{1}{b-a}$
mean: $(a+b)/2$
variance: $(b-a)^2/12$

1.4.3 Normal (Gaussian) distribution

Limit of the binomial distribution for a large number of trials. Depends on two parameters, the mean μ and the standard deviation σ. The importance of the normal distribution stems from the central limit theorem outlined below.

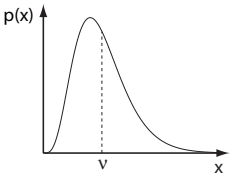


Probability density function:
$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$
mean: μ
variance: σ^2

1.4.4 Chi-square distribution

The sum of the squares of normally distributed random numbers is chi-square distributed and depends on a parameter ν that is equal to the mean. Γ is the gamma

function included in MATLAB as `gamma`.



Probability density function:
$p(x) = \frac{x^{(\nu-2)/2} e^{-x/2}}{2^{\nu/2}\Gamma(\nu/2)}$
mean: ν
variance: 2ν

1.4.5 Multivariate Gaussian distribution

We will later consider density functions of a several random variables, $x_1, ..., x_n$. Such density functions are functions in higher dimensions. An important example is the multivariate Gaussian (or Normal) distribution given by

$$p(x_1, ..., x_n) = p(\mathbf{x}) = \frac{1}{(\sqrt{(2\pi)})^n \sqrt{(}\det(\mathbf{\Sigma})} \exp(-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)). \tag{1.13}$$

This is a straight forward generalization of the one-dimensional Gaussian distribution mentioned before where the mean is now a vector, μ and the variance generalizes to a covariance matrix, $\mathbf{\Sigma} = [\mathrm{Cov}[X_i, X_j]]_{i=1,2,...,k;j=1,2,...,k}$ which must be symmetric and positive semi-definit. An example with mean $\mu = (1\ 2)^T$ and covariance $\Sigma = (1\ 0.5; 0.5\ 1)$ is shown in Fig,1.1.

1.5 Cumulative probability (density) function and the Gaussian error function

We have mainly discussed probabilities of single values as specified by the probability (density) functions. However, in many cases we want to know the probabilities of having values in a certain range. Indeed, the probability of a specific valuer of a continuous random variable is actually infinitesimally small (nearly zero), and only the probability of a range of values is finite and has a useful meaning of a probability. This integrated version of a probability density function is the probability of having a value x for the random variable X in the range of $x_1 \leq x \leq x_2$ and is given by

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} p(x)\mathrm{d}x. \tag{1.14}$$

Note that we have shortened the notation by replacing the notation $P_X(x_1 \leq X \leq x_2)$ by $P(x_1 \leq X \leq x_2)$ to simplify the following expressions. In the main text we often need to calculate the probability that a normally (or Gaussian) distributed variable has values between $x_1 = 0$ and $x_2 = y$. The probability of eqn 1.14 then becomes a function of y. This defines the **Gaussian error function**

$$\frac{1}{\sqrt{2\pi}\sigma} \int_0^y \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x = \frac{1}{2}\mathrm{erf}(\frac{y-\mu}{\sqrt{2}\sigma}). \tag{1.15}$$

The name of this function comes from the fact that this integral often occurs when calculating confidence intervals with Gaussian noise and is often abbreviated as erf.

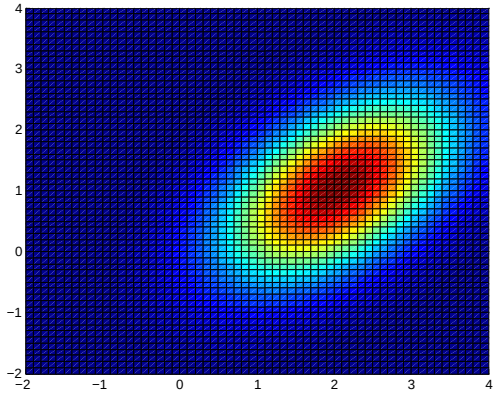


Fig. 1.1 Multivariate Gaussian with mean $\mu = (1\ 2)^T$ and covariance $\Sigma = (1\ 0.5; 0.5\ 1)$.

This Gaussian error function for normally distributed variables (Gaussian distribution with mean $\mu = 0$ and variance $\sigma = 1$) is commonly tabulated in books on statistics. Programming libraries also frequently include routines that return the values for specific arguments. In MATLAB this is implemented by the routine `erf`, and values for the inverse of the error function are returned by the routine `erfinv`.

Another special case of eqn 1.14 is when x_1 in the equation is equal to the lowest possible value of the random variable (usually $-\infty$). The integral in eqn 1.14 then corresponds to the probability that a random variable has a value smaller than a certain

value, say y. This function of y is called the **cumulative density function** (cdf),[2]

$$P^{\mathrm{cum}}(x < y) = \int_{-\infty}^{y} p(x)\mathrm{d}x, \tag{1.16}$$

which we will utilize further below.

1.6 Functions of random variables and the central limit theorem

A function of a random variable X,

$$Y = f(X), \tag{1.17}$$

is also a random variable, Y, and we often need to know what the pdf of this new random variable is. Calculating with functions of random variables is a bit different to regular functions and some care has be given in such situations. Let us illustrate how to do this with an example. Say we have an equally distributed random variable X as commonly approximated with pseudo-random number generators on a computer. The probability density function of this variable is given by

$$p_X(x) = \begin{cases} 1 & \text{for } 0 \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{1.18}$$

We are seeking the probability density function $p_Y(y)$ of the random variable

$$Y = \mathrm{e}^{-X^2}. \tag{1.19}$$

The random number Y is **not** Gaussian distributed as we might think naively. To calculate the probability density function we can employ the cumulative density function eqn 1.16 by noting that

$$P(Y \leq y) = P(\mathrm{e}^{-X^2} \leq y) = P(X \geq \sqrt{-\ln y}). \tag{1.20}$$

Thus, the cumulative probability function of Y can be calculated from the cumulative probability function of X,

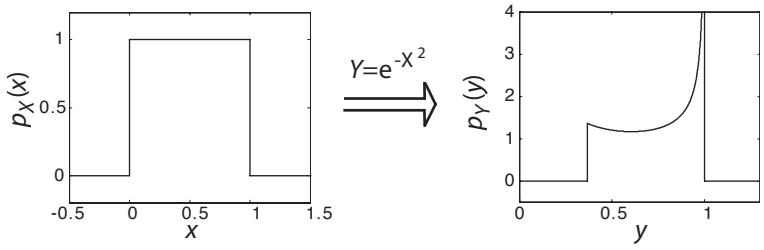
$$P(X \geq \sqrt{-\ln y}) = \begin{cases} \int_{\sqrt{-\ln y}}^{1} p_X(x)\mathrm{d}y = 1 - \sqrt{-\ln y} & \text{for } \mathrm{e}^{-1} \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{1.21}$$

The probability density function of Y is the the derivative of this function,

$$p_Y(y) = \begin{cases} 1 - \sqrt{-\ln y} & \text{for } e^{-1} \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{1.22}$$

The probability density functions of X and Y are shown below.

[2] Note that this is a probability function, not a density function.



A special function of random variables, which is of particular interest it can approximate many processes in nature, is the sum of many random variables. For example, such a sum occurs if we calculate averages from measured quantities, that is,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{1.23}$$

and we are interested in the probability density function of such random variables. This function depends, of course, on the specific density function of the random variables X_i. However, there is an important observation summarized in the **central limit theorem**. This theorem states that the average (normalized sum) of n random variables that are drawn from any distribution with mean μ and variance σ is approximately normally distributed with mean μ and variance σ/n for a sufficiently large sample size n. The approximation is, in practice, often very good also for small sample sizes. For example, the normalized sum of only seven uniformly distributed pseudo-random numbers is often used as a pseudo-random number for a normal distribution.

1.7 Measuring the difference between distributions

An important practical consideration is how to measure the similarity of difference between two density functions, say the density function p and the density function q. Note that such a measure is a matter of definition, similar to distance measures of real numbers or functions. However, a proper distance measure, d, should be zero if the items to be compared, a and b, are the same, it's value should be positive otherwise, and a distance measure should be symmetrical, meaning that $d(a,b) = d(b,a)$. The following popular measure of similarity between two density functions is not symmetric and is hence not called a distance. It is called **Kulbach–Leibler divergence** and is given by

$$d^{\mathrm{KL}}(p,q) = \int p(x)\log(\frac{p(x)}{q(x)})\mathrm{d}x \tag{1.24}$$

$$= \int p(x)\log(p(x))\mathrm{d}x - \int p(x)\log(q(x))\mathrm{d}x \tag{1.25}$$

This measure is zero if $p = q$. This measure is related to the information gain or relative entropy in information theory.

1.8 Density functions of multiple random variables

So far, we have discussed mainly probability (density) functions of single random variables. As mentioned before, we use random variables to describe data such as sensor readings in robots. Of course, we often have then more than one sensor and also other quantities that we describe by random variables at the same time. Thus, in many applications we consider multiple random variables. The quantities described by the random variables might be independent, but in many cases they are also related. Indeed, we will later talk about how to describe various types of relations. Thus, in order to talk about situations with multiple random variables, or multivariate statistics, it is useful to know basic rules. We start by illustrating these basic multivariate rules with two random variables since the generalization from there is usually quite obvious. But we will also talk about the generalization to more than two variables at the end of this section.

1.8.1 Basic definitions

We have seen that probability theory is quite handy to model data, and probability theory also considers multiple random variables. The total knowledge about the co-occurrence of specific values for two random variables X and Y is captured by the

$$\textbf{joined distribution:} \quad p(x, y) = p(X = x, Y = y). \tag{1.26}$$

This is a two dimensional functions. The two dimensions refers here to the number of variables, although a plot of this function would be a three dimensional plot. An example is shown in Fig.1.2. All the information we can have about a stochastic system is encapsulated in the joined pdf. The slice of this function, given the value of one variable, say y, is the

$$\textbf{conditional distribution}: \quad p(x|y) = p(X = x|Y = y). \tag{1.27}$$

A conditional pdf is also illustrated in Fig.1.2 If we sum over all realizations of y we get the

$$\textbf{marginal distribution}: \quad p(x) = \int p(x, y)dy. \tag{1.28}$$

If we know some functional form of the density function or have a parameterized hypothesis of this function, than we can use common statistical methods, such as maximum likelihood estimation, to estimate the parameters as in the one dimensional cases. If we do not have a parameterized hypothesis we need to use other methods, such as treating the problem as discrete and building histograms, to describe the density function of the system. Note that parameter-free estimation is more challenging with increasing dimensions. Considering a simple histogram method to estimate the joined density function where we discretize the space along every dimension into n bins. This leads to n^2 bins for a two-dimensional histogram, and n^d for a d-dimensional problem. This exponential scaling is a major challenge in practice since we need also considerable data in each bin to sufficiently estimate the probability of each bin.

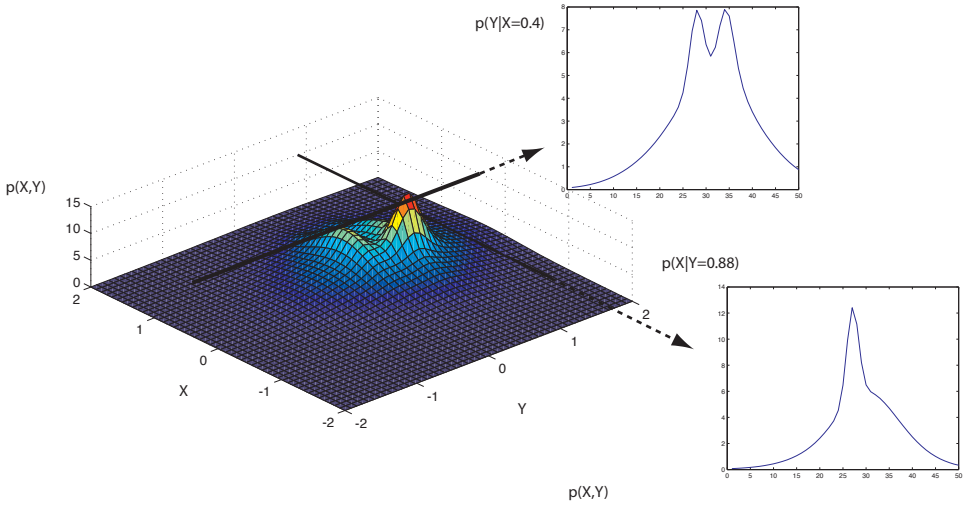


Fig. 1.2 Example of a two-dimensional probability density function (pdf) and some examples of conditional pdfs.

1.8.2 The chain rule

As mentioned before, if we know the joined distribution of some random variables we can make the most predictions of these variables. However, in practice we have often to estimate these functions, and we can often only estimate conditional density functions. A very useful rule to know is therefore how a joined distribution can be decompose into the product of a conditional and a marginal distribution,

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x), \tag{1.29}$$

which is sometimes called the **chain rule**. Note the two different ways in which we can decompose the joined distribution. This is easily generalized to n random variables by

$$\begin{aligned} p(x_1, x_2, ..., x_n) &= p(x_n|x_1, ...x_{n-1})p(x_1, ..., x_{n-1}) && (1.30)\\ &= p(x_n|x_1, ..., x_{n-1}) * ... * p(x_2|x_1) * p(x_1) && (1.31)\\ &= \Pi_{i=1}^{n} p(x_i|x_{i-1}, ...x_1) && (1.32) \end{aligned}$$

but note that there are also different decompositions possible. We will learn more about this and useful graphical representations in Chapter **??**.

Estimations of processes are greatly simplified when random variables are independent. A random variable X is independent of Y if

$$p(x|y) = p(x). \tag{1.33}$$

Using the chain rule eq.1.29, we can write this also as

$$p(x, y) = p(x)p(y), \tag{1.34}$$

that is, the joined distribution of two independent random variables is the product of their marginal distributions. Similar, we can also define conditional independence. For

example, two random variables X and Y are conditionally independent of random variable Z if

$$p(x, y|z) = p(x|z)p(y|z). \tag{1.35}$$

Note that total independence does not imply conditionally independence and visa versa, although this might hold true for some specific examples.

1.8.3 How to combine prior knowledge with new evidence: Bayes rule

One of the most common tasks we will encounter in the following is the integration of prior knowledge with new evidence. For example, we could have an estimate of the location of an agent and get new (noisy) sensory data that adds some suggestions for different locations. A similar task is the fusion of data from different sensors. The general question we have to solve is how to weight the different evidence in light of the reliability of this information. Solving this problem is easy in a probabilistic framework and is one of the main reasons that so much progress has been made in probabilistic robotics.

How prior knowledge should be combined with prior knowledge is an important question. Luckily, basically already know how to do it best in a probabilistic sense. Namely, if we divide this chain rule eq. 1.29 by $p(x)$, which is possible as long as $p(x) > 0$, we get the identity

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}, \tag{1.36}$$

which is called **Bayes theorem**. This theorem is important because it tells us how to combine a **prior** knowledge, such as the expected distribution over a random variable such as the state of a system, $p(x)$, with some evidence called the likelihood function $p(y|x)$, for example by measuring some sensors reading y when controlling the state, to get the **posterior** distribution, $p(y|x)$ from which the new estimation of state can be derived. The marginal distribution $p(y)$, which does not depend on the state X, is the proper normalization so that the left-hand side is again a probability.

Exercises

1. Use your favourite plotting program to plot a Gaussian, a uniform, and the Chi-square distribution (probability density function). Include units on the axis.
2. Explain if the random variables X and Y are independent if their marginal distribution is $p(x) = 3x^2 + log(x)$ and $p(y) = 3y^2 + log(y)$ and the joined distributions is $p(x, y) = 3x^2y^2 + log(xy)$.
3. (From Thrun, Burgard and Fox, Probabilistic Robotics) A robot uses a sensor that can measure ranges from $0m$ to $3m$. For simplicity, assume that the actual ranges are distributed uniformly in this interval. Unfortunately, the sensors can be faulty. When the sensor is faulty it constantly outputs a range below $1m$, regardless of the actual range in the sensor's measurement cone. We know that the prior probability for a sensor to be faulty is $p = 0.01$.

Suppose the robot queries its sensors N times, and every single time the measurement value is below $1m$. What is the posterior probability of a sensor fault, for $N = 1, 2, ..., 10$. Formulate the corresponding probabilistic model.

2 Programming with Matlab

This chapter is a brief introduction to programming with the Matlab programming environment. We assume thereby little programming experience, although programmers experienced in other programming languages might want to scan through this chapter. MATLAB is an interactive programming environment for scientific computing. This environment is very convenient for us for several reasons, including its interpreted execution mode, which allows fast and interactive program development, advanced graphic routines, which allow easy and versatile visualization of results, a large collection of predefined routines and algorithms, which makes it unnecessary to implement known algorithms from scratch, and the use of matrix notations, which allows a compact and efficient implementation of mathematical equations and machine learning algorithms. MATLAB stands for matrix laboratory, which emphasizes the fact that most operations are array or matrix oriented. Similar programming environments are provided by the open source systems called **Scilab** and **Octave**. The Octave system seems to emphasize syntactic compatibility with MATLAB, while Scilab is a fully fledged alternative to MATLAB with similar interactive tools. While the syntax and names of some routines in Scilab are sometimes slightly different, the distribution includes a converter for MATLAB programs. Also, the Matlab web page provides great videos to learn how to use Matlab at `http://www.mathworks.com/demos/matlab/...`
`...getting-started-with-matlab-video-tutorial.html`.

2.1 The MATLAB programming environment

MATLAB[3] is a programming environment and collection of tools to write programs, execute them, and visualize results. MATLAB has to be installed on your computer to run the programs mentioned in the manuscript. It is commercially available for many computer systems, including Windows, Mac, and UNIX systems. The MATLAB web page includes a set of brief tutorial videos, also accessible from the **demos** link from the MATLAB desktop, which are highly recommended for learning MATLAB.

As already mentioned, there are several reasons why MATLAB is easy to use and appropriate for our programming need. MATLAB is an interpreted language, which means that commands can be executed directly by an interpreter program. This makes the time-consuming compilation steps of other programming languages redundant and allows a more interactive working mode. A disadvantage of this operational mode is that the programs could be less efficient compared to compiled programs. However, there are two possible solution to this problem in case efficiency become a concern. The first is that the implementations of many MATLAB functions is very efficient

[3]MATLAB and Simulink are registered trademarks, and MATLAB Compiler is a trademark of The MathWorks, Inc.

and are themselves pre-compiled. MATLAB functions, specifically when used on whole matrices, can therefore outperform less well-designed compiled code. It is thus recommended to use matrix notations instead of explicit component-wise operations whenever possible. A second possible solutions to increase the performance is to use the MATLAB compiler to either produce compiled MATLAB code in `.mex` files or to translate MATLAB programs into compilable language such as C.

A further advantage of MATLAB is that the programming syntax supports matrix notations. This makes the code very compact and comparable to the mathematical notations used in the manuscript. MATLAB code is even useful as compact notation to describe algorithms, and it is hence useful to go through the MATLAB code in the manuscript even when not running the programs in the MATLAB environment. Furthermore, MATLAB has very powerful visualization routines, and the new versions of MATLAB include tools for documentation and publishing of codes and results. Finally, MATLAB includes implementations of many mathematical and scientific methods on which we can base our programs. For example, MATLAB includes many functions and algorithms for linear algebra and to solve systems of differential equations. Specialized collections of functions and algorithms, called a 'toolbox' in MATLAB, can be purchased in addition to the basic MATLAB package or imported from third parties, including many freely available programs and tools published by researchers. For example, the MATLAB Neural Network Toolbox incorporates functions for building and analysing standard neural networks. This toolbox covers many algorithms particularly suitable for connectionist modelling and neural network applications. A similar toolbox, called NETLAB, is freely available and contains many advanced machine learning methods. We will use some toolboxes later in this course, including the LIBSVM toolbox and the MATLAB NXT toolbox to program the Lego robots.

2.1.1 Starting a MATLAB session

Starting MATLAB opens the MATLAB desktop as shown in Fig. 2.1 for MATLAB version 7. The MATLAB desktop is comprised of several windows which can be customized or undocked (moving them into an own window). A list of these tools are available under the **desktop menu**, and includes tools such as the **command window**, **editor**, **workspace**, etc. We will use some of these tools later, but for now we only need the **MATLAB command window**. We can thus close the other windows if they are open (such as the **launch pad** or the **current directory window**); we can always get them back from the **desktop** menu. Alternatively, we can undock the command window by clicking the arrow button on the command window bar. An undocked command window is illustrated on the left in Fig. 2.2. Older versions of MATLAB start directly with a command window or simply with a MATLAB command prompt `>>` in a standard system window. The command window is our control centre for accessing the essential MATLAB functionalities.

2.1.2 Basic variables in MATLAB

The MATLAB programming environment is interactive in that all commands can be typed into the command window after the command prompt (see Fig. 2.2). The

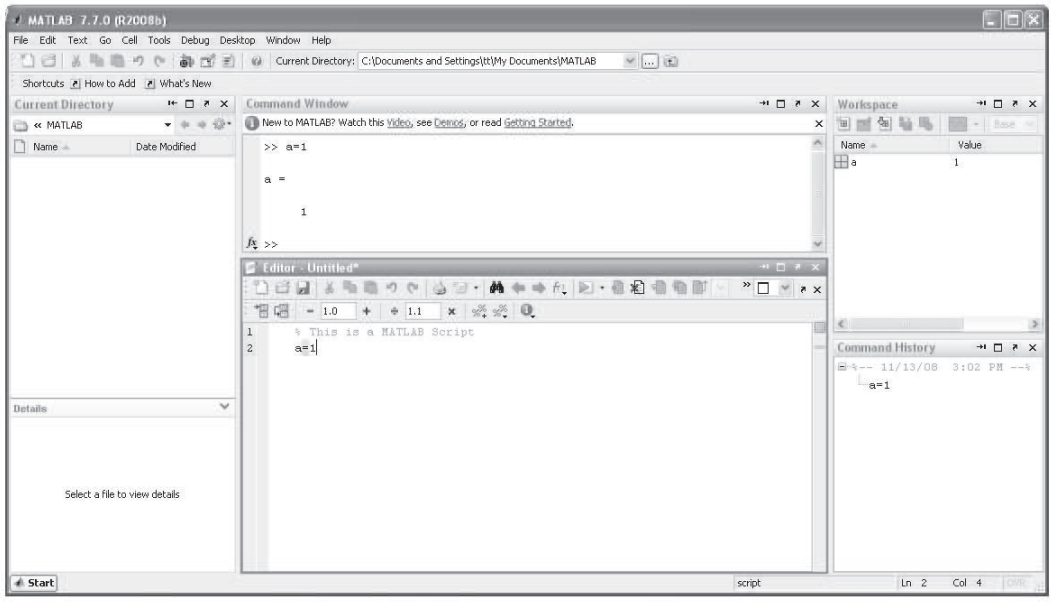


Fig. 2.1 The MATLAB *desktop window* of MATLAB Version 7.

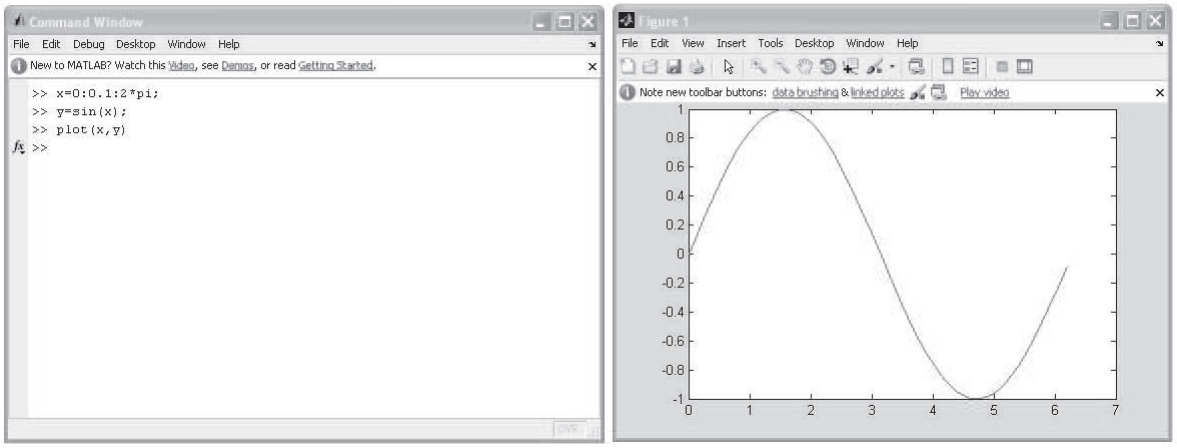


Fig. 2.2 A MATLAB *command window* (left) and a MATLAB *figure window* (right) displaying the results of the function `plot_sin` developed in the text.

commands are interpreted directly, and the result is returned to (and displayed in) the command window. For example, a variable is created and assigned a value with the = operator, such as

```
>> a=3
```

```
a =

     3
```

Ending a command with semicolon (`;`) suppresses the printing of the result on screen. It is therefore generally included in our programs unless we want to view some intermediate results. All text after a percentage sign (`%`) is not interpreted and thus treated as comment,

```
>> b='Hello World!'; % delete the semicolon to echo Hello World!
```

This example also demonstrates that the type of a variable, such as being an integer, a real number, or a string, is determined by the values assigned to the elements. This is called **dynamic typing**. Thus, variables do not have to be declared as in some other programming languages. While dynamic typing is convenient, a disadvantage is that a mistyped variable name can not be detected by the compiler. Inspecting the list of created variables is thus a useful step for debugging.

All the variables that are created by a program are kept in a buffer called **workspace**. These variable can be viewed with the command `whos` or displayed in the **workspace** window of the MATLAB desktop. For example, after declaring the variables above, the `whos` command results in the responds

```
>> whos
  Name      Size             Bytes  Class     Attributes

  a         1x1                  8  double
  b         1x12                24  char
```

It displays the name, the size, and the type (class) of the variable. The size is often useful to check the orientation of matrices as we will see below. The variables in the workspace can be used as long as MATLAB is running and as long as it is not cleared with the command `clear`. The workspace can be saved with the command `save` **filename**, which creates a file **filename**`.mat` with internal MATLAB format. The saved workspace can be reloaded into MATLAB with the command `load` **filename**. The load command can also be used to import data in ASCII format. The workspace is very convenient as we can run a program within a MATLAB session and can then work interactively with the results, for example, to plot some of the generated data.

Variables in MATLAB are generally matrices (or data arrays), which is very convenient for most of our purposes. Matrices include scalars (1×1 matrix) and vectors ($1 \times N$ matrix) as special cases. Values can be assigned to matrix elements in several ways. The most basic one is using square brackets and separating rows by a semicolon within the square brackets, for example (see Fig. 2.2),

```
>> a=[1 2 3; 4 5 6; 7 8 9]

a =

     1     2     3
     4     5     6
     7     8     9
```

A vector of elements with consecutive values can be assigned by column operators like

```
>> v=0:2:4

v =

     0     2     4
```

Furthermore, the MATLAB desktop includes an **array editor**, and data in ASCII files can be assigned to matrices when loaded into MATLAB. Also, MATLAB functions often return matrices which can be used to assign values to matrix elements. For example, a uniformly distributed random 3×3 matrix can be generated with the command

```
>> b=rand(3)

b =

    0.9501    0.4860    0.4565
    0.2311    0.8913    0.0185
    0.6068    0.7621    0.8214
```

The multiplication of two matrices, following the matrix multiplication rules, can be done in MATLAB by typing

```
>> c=a*b

c =

    3.2329    4.5549    2.9577
    8.5973   10.9730    6.8468
   13.9616   17.3911   10.7360
```

This is equivalent to

```
c=zeros(3);
for i=1:3
   for j=1:3
      for k=1:3
         c(i,j)=c(i,j)+a(i,k)*b(k,j);
      end
   end
end
```

which is the common way of writing matrix multiplications in other programming languages. Formulating operations on whole matrices, rather than on the individual components separately, is not only more convenient and clear, but can enhance the programs performance considerably. Whenever possible, operations on whole matrices should be used. This is likely to be the major change in your programming style when converting from another programming language to MATLAB. The performance disadvantage of an interpreted language is often negligible when using operations on

whole matrices.

The transpose operation of a matrix changes columns to rows. Thus, a row vector such as v can be changed to a column vector with the MATLAB transpose operator (’),

```
>> v'

ans =

     0
     2
     4
```

which can then be used in a matrix-vector multiplication like

```
>> a*v'

ans =

    16
    34
    52
```

The inconsistent operation a*v does produce an error,

```
>> a*v
??? Error using ==> mtimes
Inner matrix dimensions must agree.
```

Component-wise operations in matrix multiplications (*), divisions (/) and potentiation ^ are indicated with a dot modifier such as

```
>> v.^2

ans =

     0     4    16
```

The most common operators and basic programming constructs in MATLAB are similar to those in other programming languages and are listed in Table 2.1.

2.1.3 Control flow and conditional operations

Besides the assignments of values to variables, and the availability of basic data structures such as arrays, a programming language needs a few basic operations for building loops and for controlling the flow of a program with conditional statements (see Table 2.1). For example, the **for loop** can be used to create the elements of the vector v above, such as

```
>> for i=1:3; v(i)=2*(i-1); end
>> v

v =
```

Table 2.1 Basic programming contracts in MATLAB.

Programming construct	Command	Syntax
Assignment	=	a=b
Arithmetic operations	add	a+b
	multiplication	a*b (matrix), a.*b (element-wise)
	division	a/b (matrix), a./b (element-wise)
	power	a^b (matrix), a.^b (element-wise)
Relational operators	equal	a==b
	not equal	a~=b
	less than	a<b
Logical operators	AND	a & b
	OR	a||b
Loop	for	for **index=start:increment:end** statement end
	while	while **expression** statement end
Conditional command	if statement	if **logical expressions** statement elseif **logical expressions** statement else statement end
Function		function [x,y,...]=**name**(a,b,...)

```
     0     2     4
```

Table 2.1 lists, in addition, the syntax of a while loop. An example of a conditional statement within a loop is

```
>> for i=1:10; if i>4 & i<=7; v2(i)=1; end; end
>> v2

v2 =

     0     0     0     0     1     1     1
```

In this loop, the statement `v2(i)=1` is only executed when the loop index is larger than 4 and less or equal to 7. Thus, when `i=5`, the array `v2` with 5 elements is created, and since only the elements `v2(5)` is set to 1, the previous elements are set to 0 by default. The loop adds then the two element `v2(6)` and `v2(7)`. Such a vector can also be created by assigning the values 1 to a specified range of indices,

```
>> v3(4:7)=1

v3 =
```

```
     0     0     0     1     1     1     1
```

A 1×7 array is thereby created with elements set to 0, and only the specified elements are overwritten with the new value 1. Another method to write compact loops in MATLAB is to use vectors as index specifiers. For example, another way to create a vector with values such as `v2` or `v3` is

```
>> i=1:10

i =

     1     2     3     4     5     6     7     8     9    10

>> v4(i>4 & i<=7)=1

v4 =

     0     0     0     0     1     1     1
```

2.1.4 Creating MATLAB programs

If we want to repeat a series of commands, it is convenient to write this list of commands into an ASCII file with extension '`.m`'. Any ASCII editor (for example; WordPad, Emacs, etc.) can be used. The MATLAB package contains an editor that has the advantage of colouring the content of MATLAB programs for better readability and also provides direct links to other MATLAB tools. The list of commands in the ASCII file (e.g. **prog1**`.m`) is called a **script** in MATLAB and makes up a MATLAB program. This program can be executed with a run button in the MATLAB editor or by calling the name of the file within the command window (for example, by typing **prog1**). We assumed here that the program file is in the current directory of the MATLAB session or in one of the search paths that can be specified in MATLAB. The MATLAB desktop includes a 'current directory' window (see desktop menu). Some older MATLAB versions have instead a 'path browser'. Alternatively, one can specify absolute path when calling a program, or change the current directories with UNIX-style commands such as `cd` in the command window (see Fig. 2.3).

Functions are another way to encapsulate code. They have the additional advantage that they can be pre-compiled with the MATLAB Compiler$^{\mathrm{TM}}$ available from MathWorks, Inc. Functions are kept in files with extension '`.m`' which start with the command line like

```
function y=f(a,b)
```

where the variables `a` and `b` are passed to the function and `y` contains the values returned by the function. The return values can be assigned to a variable in the calling MATLAB

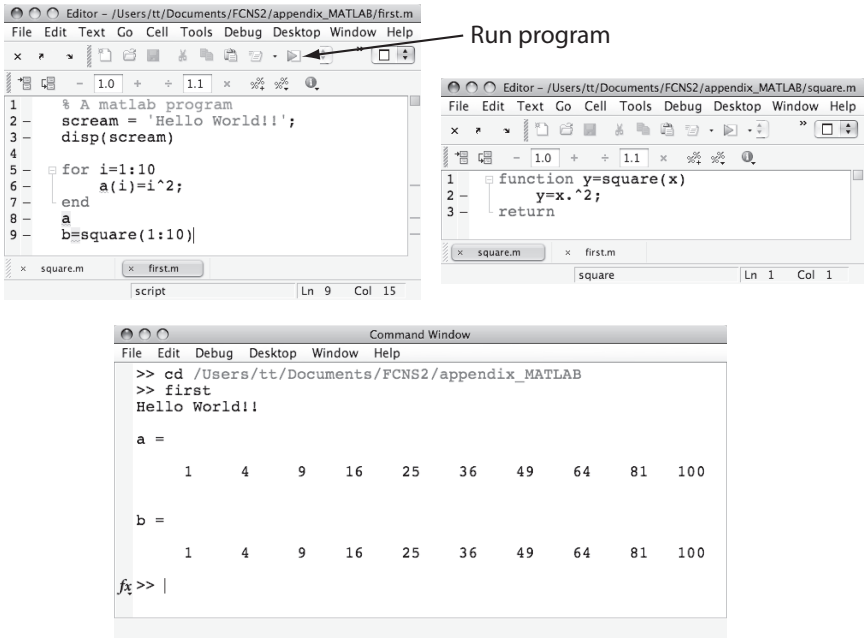


Fig. 2.3 Two editor windows and a command window.

script and added to the workspace. All other internal variables of the functions are local to the function and will not appear in the workspace of the calling script.

MATLAB has a rich collection of predefined functions implementing many algorithms, mathematical constructs, and advanced graphic handling, as well as general information and help functions. You can always search for some keywords using the useful command `lookfor` followed by the keyword (search term). This command lists all the names of the functions that include the keywords in a short description in the function file within the first **comment lines** after the function declaration in the function file. The command `help`, followed by the function name, displays the first block of comment lines in a function file. This description of functions is usually sufficient to know how to use a function. A list of some frequently used functions is listed in Table 2.1.4.

2.1.5 Graphics

MATLAB is a great tool for producing scientific graphics. We want to illustrate this by writing our first program in MATLAB: calculating and plotting the sine function. The program is

```
x=0:0.1:2*pi;
y=sin(x);
plot(x,y)
```

The first line assigns elements to a vector $\mathbf{x}$ starting with $x(1) = 0$ and incrementing the value of each further component by 0.1 until the value 2π is reached (the variable

Name	Brief description	Name	Brief description
abs	absolute functions	mod	modulus function
axis	sets axis limits	num2str	converts number to string
bar	produces bar plot	ode45	ordinary differential equation solver
ceil	round to larger interger	ones	produces matrix with unit elements
colormap	colour matrix for surface plots	plot	plot lines graphs
cos	cosine function	plot3	plot 3-dimensional graphs
diag	diagonal elements of a matrix	prod	product of elements
disp	display in command window	rand	uniformly distributed random variable
errorbar	plot with error bars	randn	normally distributed random variable
exp	exponential function	randperm	random permutations
fft	fast Fourier transform	reshape	reshaping a matrix
find	index of non-zero elements	set	sets values of parameters in structure
floor	round to smaller integer	sign	sign function
hist	produces histogram	sin	sine function
int2str	converts integer to string	sqrt	square root function
isempty	true if array is empty	std	calculates standard deviation
length	length of a vector	subplot	figure with multiple subfigures
log	logarithmic function	sum	sum of elements
lsqcurevfit	least mean square curve	surf	surface plot
	fitting (statistics toolbox)	title	writes title on plot
max	maximum value and index	view	set viewing angle of 3D plot
mix	minimum value and index	xlabel	label on x-axis of a plot
mean	calculates mean	ylabel	label on y-axis of a plot
meshgrid	creates matrix to plot grid	zeros	creates matrix of zero elements

Table 2.2 MATLAB functions used in this course. The MATLAB command `help cmd`, where `cmd` is any of the functions listed here, provides more detailed explanations.

`pi` has the appropriate value in MATLAB). The last element is $x(63) = 6.2$. The second line calls the MATLAB function `sin` with the vector `x` and assigns the results to a vector `y`. The third line calls a MATLAB plotting routine. You can type these lines into an ASCII file that you can name `plot_sin.m`. The code can be executed by typing `plot_sin` as illustrated in the **command window** in Fig. 2.2, provided that the MATLAB session points to the folder in which you placed the code. The execution of this program starts a **figure window** with the plot of the sine function as illustrated on the right in Fig. 2.2.

The appearance of a plot can easily be changed by changing the attributes of the plot. There are several functions that help in performing this task, for example, the function `axis` that can be used to set the limits of the axis. New versions of MATLAB provide window-based interfaces to the attributes of the plot. However, there are also two basic commands, `get` and `set`, that we find useful. The command `get(gca)` returns a list with the axis properties currently in effect. This command is useful for finding out what properties exist. The variable `gca` (get current axis) is the **axis handle**, which is a variable that points to a memory location where all the attribute variables are kept. The attributes of the axis can be changed with the `set` command. For example, if we want to change the size of the labels we can type `set(gca,'fontsize',18)`.

There is also a handle for the current figure `gcf` that can be used to get and set other attributes of the figure. MATLAB provides many routines to produce various special forms of plots including plots with error-bars, histograms, three-dimensional graphics, and multi-plot figures.

2.2 A first project: modelling the world

Suppose there is a simple world with a creature that can be in three distinct states, sleep (state value 1), eat (state value 2), and study (state value 3). An agent, which is a device that can sense environmental states and can generate actions, is observing this creature with poor sensors, which add white (Gaussian) noise to the true state. Our aim is to build a model of the behaviour of the creature which can be used by the agent to observe the states of the creature with some accuracy despite the limited sensors. For this exercise, the function `creature_state()` is available on the course page on the web. This function returns the current state of the creature. Try to create an agent program that predicts the current state of the creature. In the following we discuss some simple approches.

A simulation program that implements a specific agent a with simple world model (a model of the creature), which also evaluates the accuracy of the model, is given in Table 2.3. This program, also available on the web, is provided in file `main.m`. This program can be downloaded into the working directory of MATLAB and executed by typing `main` into the command window, or by opening the file in the MATLAB editor and starting it from there by pressing the icon with the green triangle. The program reports the percentage of correct perceptions of the creature's state.

Line 1 of the program uses a comment indicator (%) to outline the purpose of the program. Line 2 clears the workspace to erase all eventual existing variables, and sets a counter for the number of correct perceptions to zero. Line 4 starts a loop over 1000 trials. In each trial, a creature state is pulled by calling the function `creature_state()` and recording this state value in variable `x`. The sensory state `s` is then calculated by adding a random number to this value. The value of the random number is generated from a normal distribution, a Gaussian distribution with mean zero and unit variance, with the MATLAB function `randn()`.

We are now ready to build a model for the agent to interpret the sensory state. In the example shown, this model is given in Lines 8–12. This model assumes that a sensory value below 1.5 corresponds to the state of a sleeping creature (Line 9), a sensory value between 1.5 and 2.5 corresponds to the creature eating (Line 10), and a higher value corresponds to the creature studying (Line 11). Note that we made several assumptions by defining this model, which might be unreasonable in real-world applications. For example, we used our knowledge that there are three states with ideal values of 1, 2, and 3 to build the model for the agent. Furthermore, we used the knowledge that the sensors are adding independent noise to these states in order to come up with the decision boundaries. The major challenge for real agents is to build models without this explicit knowledge. When running the program we find that a little bit over 50% of the cases are correctly perceived by the agent. While this is a good start, one could do better. Try some of your own ideas . . .

Table 2.3 Program main.m

```
% Project 1: simulation of agent which models simple creature
clear; correct=0;

for trial=1:1000
    x=creature_state();
    s=x+randn();

    %% perception model
    if (s<1.5) x_predict=1;
    elseif (s<2.5) x_predict=2;
    else x_predict=3;
    end

    %% calculate accuracy
    if (x==x_predict) correct=correct+1; end
 end

disp(['percentage correct: ',num2str(correct/1000)]);
```

. . . Did you succeed in getting better results? It is certainly not easy to guess some better model, and it is time to inspect the data more carefully. For example, we can plot the number of times each state occurs. For this we can write a loop to record the states in a vector,

```
>> for i=1:1000; a(i)=creature_state(); end
```

and then plot a histogram with the MATLAB function `hist()`,

```
>> hist(a)
```

The result is shown in Fig. 2.4. This histogram shows that not all states are equally likely as we implicitly assumed in the above agent model. The third state is indeed much less likely. We could use this knowledge in a modified model in which we predict that the agent is sleeping for sensory states less than 1.5 and is eating otherwise. This modified model, which completely ignores study states, predicts around 65% of the states correctly. Many machine learning methods suffer from such 'explaining away' solutions for imbalanced data, as further discussed in Chapter **??**.

It is important to recognize that 100% accuracy is not achievable with the inherent limitations of the sensors. However, higher recognition rates could be achieved with better world (creature + sensor) models. The main question is how to find such a model. We certainly should use observed data in a better way. For example, we could use several observations to estimate how many states are produced by function `creature_state()` and their relative frequency. Such parameter estimation is a basic form of learning from data. Many models in science take such an approach by proposing a parametric model and estimating parameters from the data by model fitting. The main challenge with this approach is how complex we should make the model. It is much easier to fit a more complex model with many parameters to example data, but the

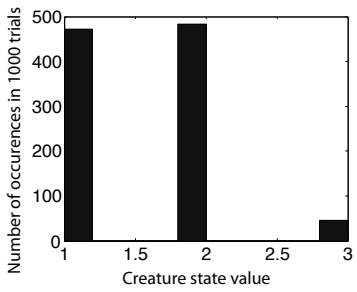


Fig. 2.4 The MATLAB *desktop window* histogram of states produced by function `creature_state()` from 1000 trials.

increased flexibility decreases the prediction ability of such models. Much progress has been made in machine learning by considering such questions, but those approaches only work well in limited worlds, certainly much more restricted than the world we live in. More powerful methods can be expected by learning how the brain solves such problems.

2.3 Alternative programming environments: Octave and Scilab

We briefly mention here two programming environments that are very similar to Matlab and that can, with certain restrictions, execute Matlab scripts. Both of these programming systems are open source environments and have general public licenses for non-commercial use.

The programming environment called **Octave** is freely available under the GNU general public license. Octave is available through links at http://www.gnu.org/software/octave/. The installation requires the additional installation of a graphics package, such as gnuplot or Java graphics. Some distributions contain the SciTE editor which can be used in this environment. An example of the environment is shown in Fig. 2.5

Scilab is another scientific programming environment similar to MATLAB. This software package is freely available under the CeCILL software license, a license compatible to the GNU general public license. It is developed by the Scilab consortium, initiated by the French research centre INRIA. The Scilab package includes a MATLAB import facility that can be used to translate MATLAB programs to Scilab. A screen shot of the Scilab environment is shown in Fig. 2.6. A Scilab script can be run from the execute menu in the editor, or by calling `exec("filename.sce")`.

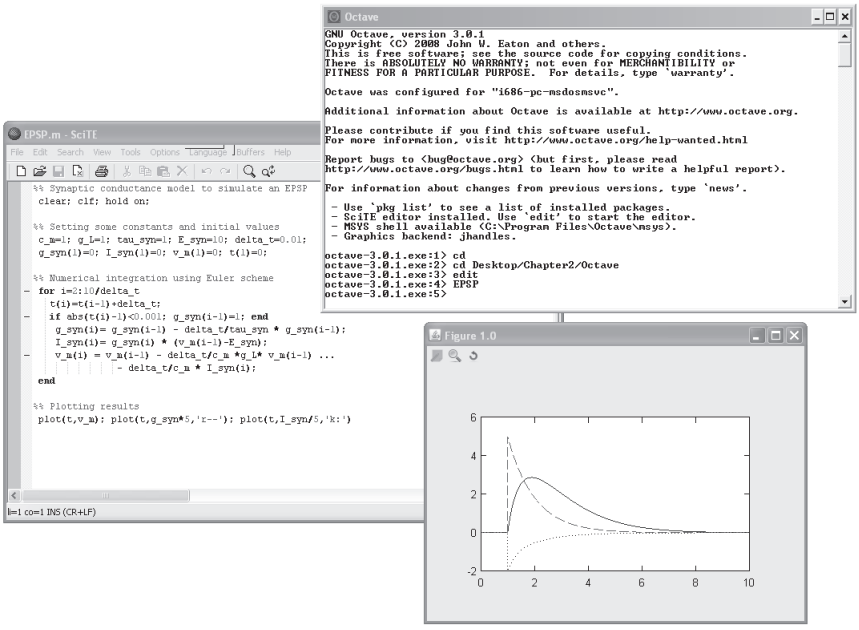


Fig. 2.5 The Octave programming environment with the main console, and editor called *SciTE*, and a graphics window.

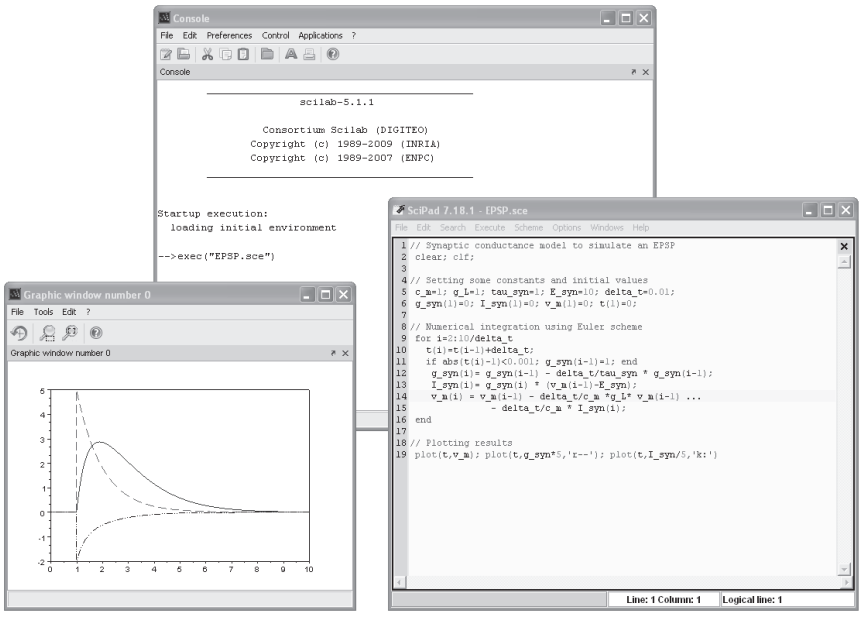


Fig. 2.6 The Scilab programming environment with *console*, and editor called *SciPad*, and a graphics window.