3 Probability theory and motion/sensor models

A major milestone for modern approaches to machine learning and robotics is the acknowledgement of 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 numbers**. Random numbers are a wonderful mathematical construct to describe uncertainty. While a specific regular number has only one specific value, a random number will have different values every time we query it. Each query is drawing an example from a probability distributions that governs this variable. For example, consider 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 variable batteries, or since we move the sensor unintentionally away from the surface, it is more than likely that we get different readings over time. Consequently, variables that have to be estimated from sensors, such as the pose of a robot, are fundamentally random numbers.

A common misconception about randomness is that one can not predict anything for random numbers. But even random numbers have most commonly values that are 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 about how often a certain number will appear when drawing many examples. We might even be able to state how confident we are that a specific number occurs, or, in other words, how uncertain a specific value might be or how it might vary when drawing several examples. The complete knowledge of a random number, that is, how likely each value is for a random number x, is captured by the **probability density function** $p(x)$. 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 numbers, 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.

3.1 Random numbers and their probability density function

3.1.1 How to describe uncertainty

Probability theory is the theory of **uncertainty** that uses the construct of **random numbers** as mathematical formalism. We denoted such numbers, or more precisely random variables when we write symbols to represent them, by capital letters to distinguish them from regular variables written in lower case. A random number, 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 (sample) of a random number is still meaningful and might influence subsequent processes in a specific (even deterministic) way.

Since a random number can change values every time it is inspected, it is useful to describe properties of a distribution 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 in the 'frequentist' interpretation of random numbers by the mathematical construct of a **probability**. A slightly different interpretation of a random numbers is that it describes the uncertainty that comes with each drawing of a random number. This 'Bayesian' interpretation is useful as we would then be comfortable applying such constructs even to events that we can not repeat easily. This view is sometimes contrasted with a 'Frequnetist' interpretation that stresses that such theories only make sense when sampling many times. For most of what we discuss in the following we mainly apply this formalism, and there is little need to twelve into a philosophical debate. There are even other formulations and formalization, such as **fuzzy systems**, which capture many aspects of the following discussions.

We can formalize the idea of expressing probabilities of drawing specific values for random number with some compact notations. We speak of a **discrete random number** in the case of discrete numbers for the possible values of a random number. A **continuous random number** is a random number 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 (mass) function**,

$$P(x) = P(X = x) \tag{3.1}$$

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 number. The value of $P(x)$ gives the fraction of the times we get a value x for the random number X if we draw many examples of the random number.[3] From this definition it follows that the frequency of having any of the possible values is equal to one, which is the normalization condition

$$\sum_x P(x) = 1. \tag{3.2}$$

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

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 $P(x) = p(x)\mathrm{d}x$, where the lower case function $p(x)$ is the **probability density function** (pdf). To be precise, this density function is again a short hand notation of

$$p(x) = p(X = x) \tag{3.3}$$

in analogy with eqn 3.1, but the sum in eqn 3.2 then becomes an integral, and normalization condition for a continuous random number is

$$\int_x p(x)\mathrm{d}x = 1. \tag{3.4}$$

There is of course an infinite number of values for a continuous variable in some interval. Therefore, only the probability of getting a number in a certain interval could have a meaningful finite value such as

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} p(x)\mathrm{d}x. \tag{3.5}$$

Another useful description of a continuous random variable is the **cumulative density function** (cdf)

$$P_c(x) = P(-\infty < X < x) = \int_{-\infty}^{x} p(x')\mathrm{d}x'. \tag{3.6}$$

We will formulate the rest of this section in terms of continuous random numbers. 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 a continuous formulation of discrete random numbers with the mathematical construct of a δ-function. Thus, the differences between continuous and discrete random numbers are mainly technical and will hopefully not distract from the general ideas.

An important basic examples of a discrete random number is a **Bernoulli random number**. A Bernoulli random number is a binary number drawn 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: $p(1 - p)$

Another important discrete distribution is the **multinomial distribution**, which is the distribution of outcomes in n trials that have k possible outcomes. The probability

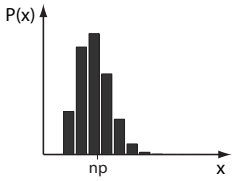
of each outcome is thereby p_i.

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

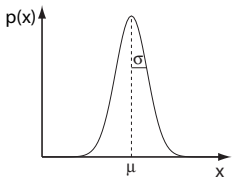
$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \tag{3.7}$$

and is given by the MATLAB function `nchoosek`.



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

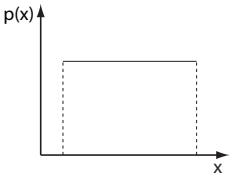
A very important example of a continuous random number is a **Gaussian or Normal distributed random number**. The Normal or Gaussian distribution describes a continuous random number with a single bell shaped peak in the distribution as shown below. The pdf depends on two parameters, the mean μ and the standard deviation σ. The importance of the normal distribution stems from the central limit theorem outlined below. This theorem captures an interesting fact about sums of random numbers, namely that the sum of many random numbers is Gaussian distributed. Formally, this is only correctly true if the random numbers are independent and drawn from the same (but arbitrary) distribution, and also that an infinite number of such variables is considered. But the importance in practice is that even small sums of random numbers with different underlying pdfs have often a distribution that is well approximated by a Gaussian.



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

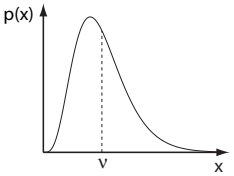
Another important distribution that we encounter in the following is the **uniform distribution**. This is simply the distribution where random numbers are equally likely in an interval $a \leq x \leq b$. Pseudo-random variables with this distribution are often

generated by routines in many programming languages.



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

Finally one example of a continuous distribution with an unsymmetric shape. 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`.



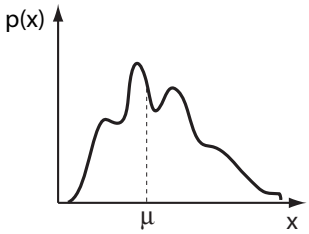
Probability density function:
$p(x) = \frac{x^{(\nu-2)/2}e^{-x/2}}{2^{\nu/2}\Gamma(\nu/2)}$
mean: ν
variance: 2ν

There are many more examples of density functions with specific names, but it is equally important to realize that there is an infinite number of possible pdfs. In general we do not know the pdf of random variables encounter in our environment or that describe our robot. Estimating the density function is the main challenge of machine learning.

3.1.2 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 probability 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, $p(x)$, 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 3.6. 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. However, sometimes it is sufficient to know at least some basic characteristics of the functions. 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. A more common quantity to know is the expected arithmetic average of those numbers, which is called the **mean**, **expected value**, or **expectation value** of the distribution, defined by

$$\mu = \int_{-\infty}^{\infty} xp(x)\mathrm{d}x. \tag{3.8}$$

In the discrete case, this formula corresponds to the formula of calculating an arithmetic average, where we add up all the different numbers together with their frequency. Formally, we need to distinguish between a quantity calculated from random numbers and quantities calculated from the pdfs. If we treat the pdf as fundamental, then the arithmetic average is like an estimation of the mean. This is usually how it is viewed. 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. While this is mainly a philosophical debate, we try to be consistent in calling the quantities derived from data 'estimates' of the quantities defined through pdfs.

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

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)\mathrm{d}x. \tag{3.9}$$

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

$$m^n = \int_{-\infty}^{\infty} (x-\mu)^n p(x)\mathrm{d}x. \tag{3.10}$$

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

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)\mathrm{d}x. \tag{3.11}$$

Higher moments specify further characteristics such as the kurtosis and skewness of the distribution. Moments higher than this have not been given explicit names. Knowing all moments of a distribution is equivalent in knowing the distribution precisely, and knowing a pdf is equivalent in 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**,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{3.12}$$

and the variance either from the **biased sample variance**,

$$s_1^2 = \frac{1}{n}\sum_{i=1}^{n}(\bar{x} - x_i)^2, \tag{3.13}$$

or the **unbiased sample variance**

$$s_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{x} - x_i)^2. \tag{3.14}$$

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. Knowing all moments uniquely specifies a pdf.

3.1.3 Functions of random variables and the central limit theorem

A function of a random variable X,

$$Y = f(X), \tag{3.15}$$

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

$$p(x) = \begin{cases} 1 & \text{for } 0 \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3.16}$$

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

$$Y = \mathrm{e}^{-X^2}. \tag{3.17}$$

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 3.6 by noting that

$$P(Y \leq y) = P(\mathrm{e}^{-X^2} \leq y) = P(X \geq \sqrt{-\ln y}). \tag{3.18}$$

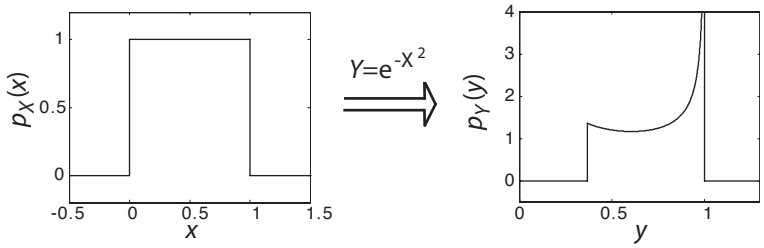
Thus, the cumulative probability function of Y can be calculated from the cumulative probability function of X,

$$P(X \geq \sqrt{-\ln y}) = \begin{cases} \int_{\sqrt{-\ln y}}^{1} p(x)\mathrm{d}y = 1 - \sqrt{-\ln y} & \text{for } \mathrm{e}^{-1} \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3.19}$$

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

$$p(y) = \begin{cases} 1 - \sqrt{-\ln y} & \text{for } e^{-1} \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3.20}$$

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



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,

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i, \tag{3.21}$$

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.

3.1.4 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, its 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

$$d^{\mathrm{KL}}(p, q) = \int p(x) \log(\frac{p(x)}{q(x)}) \mathrm{d}x \tag{3.22}$$

$$= \int p(x) \log(p(x)) \mathrm{d}x - \int p(x) \log(q(x)) \mathrm{d}x \tag{3.23}$$

This measure is zero if $p = q$ since $log(1) = 0$. This measure is related to the information gain or relative entropy in information theory.

3.2 Density functions of multiple random numbers

3.2.1 Multivariate distributions

So far, we have discussed mainly probability (density) functions of single random numbers. As mentioned before, we use random numbers 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 numbers at the same time. Thus, in many applications we consider multiple random numbers. The quantities described by the random numbers 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 numbers, or multivariate statistics, it is useful to know basic rules. We start by illustrating these basic multivariate rules with two random numbers 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.

An example of a multivariate density function over several random numbers, $x_1, ..., x_n$ is the multivariate Gaussian (or Normal) distribution,

$$p(x_1, ..., x_n) = p(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^n \sqrt{\det(\mathbf{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right). \quad (3.24)$$

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,3.1.

In robotics systems as well as in many realistic machine learning problems we have systems that depend on many random numbers. An efficient way to argue in such multivariate cases is hence crucial for real world applications. We will come back to this are when considering Bayesian models.

3.2.2 Joined, conditional and marginal distributions

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

$$\textbf{joined distribution:} \quad p(x, y) = p(X = x, Y = y). \quad (3.25)$$

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.3.2. 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

$$\textbf{conditional distribution}: \quad p(x|y) = p(X = x|Y = y). \quad (3.26)$$

A conditional pdf is also illustrated in Fig.3.2 If we sum over all realizations of y we get the

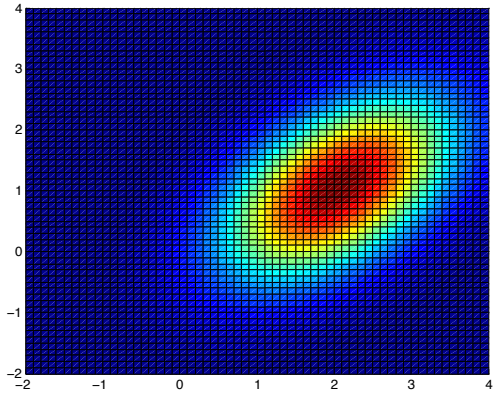


Fig. 3.1 Multivariate Gaussian with mean $\mu = (1\ 2)^T$ and covariance $\Sigma = (1\ 0.5; 0.5\ 1)$.

$$\textbf{marginal distribution}: \quad p(x) = \int p(x, y)dy, \tag{3.27}$$

which is sometimes called the **sum rule** or **marginalization**.

In some cases we might have or guess a functional form of the density function which typically has parameters that we need to estimate from data. With a **parameterized hypothesis**, 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.

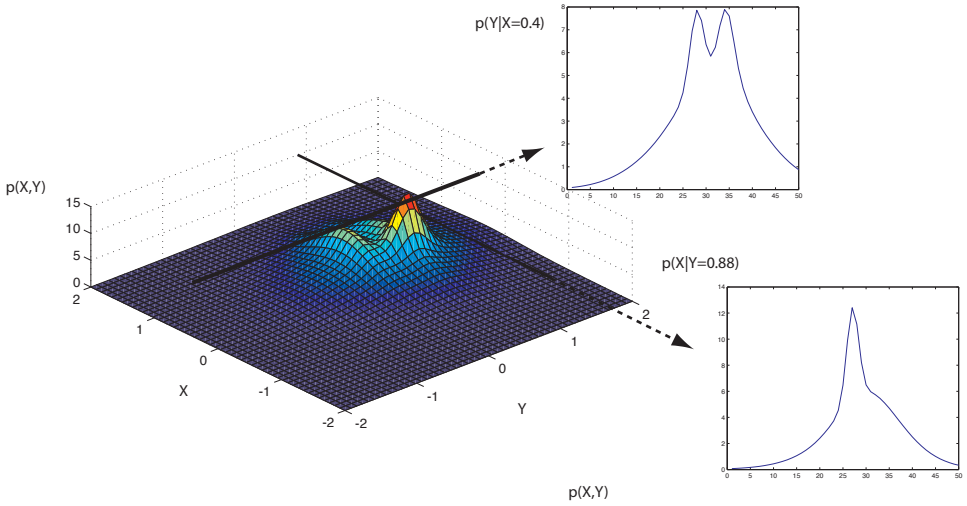


Fig. 3.2 Example of a two-dimensional probability density function (pdf) and some examples of conditional pdfs.

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.

3.2.3 The chain rule

As mentioned before, if we know the joined distribution of some random numbers 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,

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x), \tag{3.28}$$

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

$$\begin{aligned} p(x_1, x_2, ..., x_n) &= p(x_n|x_1, ...x_{n-1})p(x_1, ..., x_{n-1}) && (3.29)\\ &= p(x_n|x_1, ..., x_{n-1}) * ... * p(x_2|x_1) * p(x_1) && (3.30)\\ &= \Pi_{i=1}^{n} p(x_i|x_{i-1}, ...x_1) && (3.31) \end{aligned}$$

but note that there are also different decompositions possible. We will learn more about this and useful graphical representations in Chapter 5.

Estimations of processes are greatly simplified when random numbers are independent. A random number X is independent of Y if

$$p(x|y) = p(x). \tag{3.32}$$

Using the chain rule eq.3.28, we can write this also as

$$p(x, y) = p(x)p(y), \tag{3.33}$$

that is, the joined distribution of two independent random numbers is the product of their marginal distributions. Similar, we can also define conditional independence. For example, two random numbers X and Y are conditionally independent of random number Z if

$$p(x, y|z) = p(x|z)p(y|z). \tag{3.34}$$

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

3.2.4 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. 3.28 by $p(x)$, which is possible as long as $p(x) > 0$, we get the identity which is called **Bayes theorem**

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}, \tag{3.35}$$

This identity is important because it tells us what the probability of a given state is give a measurement of y, that is, $p(x|y)$. Prior to the measurement our believe of a state x is given by $p(x)$. Furthermore, the likelihood of having a measurement y is given by the likelihood functions $p(y|x)$. Bayes theorem tells us that we should weight there likelihood of measurement y with the prior probability of being in the corresponding state, and then normalize this number by the marginal distributions

$$p(y) = \int p(y|x)p(x)\mathrm{d}x \tag{3.36}$$

The marginal probability of y, $p(y)$, does not depend on the state x. However, estimating the denominator $p(y)$ is often the most labor intensive part of applying Bayes theorem since we often mainly mainly estimate $p(y|x)$ so that we have keep a running average

of this density over all possible states to estimate this marginal distribution. It is sometimes useful to remember Bayes rule in the form

$$p(x|y) = \frac{p(y|x)p(x)}{\int p(y|x)p(x)dx}. \tag{3.37}$$

where we explicitly replaced the marginal distribution of y. We will see below that Bayes rule (eq. 3.37) together with the rule of total probability (eq. 3.3) are the basic rules that allow us to argue (making statistical inference) in probabilistic systems.

3.2.5 Matlab support for statistics

Random numbers and statistics are a common tool in science so that support in this area is not surprising. We are mainly concerned here with using random numbers and probability theory rather than statistical method for hypothesis testing. Here we only discuss producing the some basic random numbers.

Matlab has some support for random numbers, in particular to generate the most frequently used ones. For example, the function

```
rand()
```

generates uniformly distributed random numbers between 0 and 1. A normal distributed random number can be generated with

```
randn()
```

An 10×2 array of normal distributed numbers can be generated with

```
randn(10,2)
```

A 2×5 matrix of uniformly distributed integer random numbers between 1 and maxn (inclusive) can be generated with

```
randi(maxn,2,5)
```

The statistics toolbox has much more support for random variables, including several more random number generators for different distributions, support for cumulative distributions, and some machine learning tools that we discuss later.

Exercises

1. Plot a histogram of random numbers drawn from the Chi-square distribution and the Trappenberg distribution. The Trappenberg distribution is given by

$$p(x) = \begin{cases} a_n || \sin(x)|| & \text{for } 0 < x < n\pi/2 \\ 0 & \text{otherwise} \end{cases} \tag{3.38}$$

   for $n = 5$. What is the mean, variance, and skewness of these distributions?
2. Explain if the random numbers X and Y are independent if their marginal distribution is $p(x) = x + 3log(x)$ and $p(y) = 3ylog(y)$, and the joined distributions is $p(x, y) = xylog(x) + 3ylog(xy)$.
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.

4. Given are four Bernoulli distributed random numbers X_1, X_2, X_3 and Y. The conditional probability of random numbers X_i on Y is given by $p(x_i|y) = 0.2$ and are conditionally independent of each other given Y. The marginal probability of Y is $p(y) = 0.3$. What is the probability of Y given X_1 =true and X_2 =true and X_3 =false ?

3.3 Probabilistic sensor models

Having acknowledged that sensors are noisy, we now turn to their corresponding probabilistic description that will be used later in this course. A **sensor model** describes the likelihood of a sensor value, which we denote here with x, given a certain reading of the sensor, denoted by Z. That is, a probabilistic sensor model describes

$$\textbf{Sensor model: } p(x|Z). \tag{3.39}$$

We make here the implicit assumption that this measurement model does not depend on the history of previous states. To be more specific, let us make a specific sensor model for the ultrasonic sensor. To investigate how this sensor responds to an obstacle at different distances we need a tape measure to measure the actual distance of the ultrasonic sensor from the obstacle. We then read the sensor for different distances from the obstacle and repeat this several times. Examples of such measurement for two different ultrasonic sensors are shown in Fig. 3.3. Other testing scenarios or shown in 3.4.

The figures reveal some interesting observations. While the return value is engineered to return the distance to an obstacle in cm, these tests show that the distance to the obstacle in our measurements are always overestimated. Furthermore, the error is not the same for all distances Interestingly there is an intermediate area where the error is largest and the error gets smaller for large distances. The precise response of the sensor depends on many factors such as the material of the obstacle, the orientation of it's reflecting surface, and the general conditions of the surroundings. Dealing with such uncertainties are the major focus of our investigations later in this course.

For now let us just concentrate on variations even within the specific environment in which we made these measurements. The measurements were repeated several times, and since the system responds only with discrete values, the number times a specific number was reported in indicated by the size of the circle representing the measurement. It is clear that there is some distribution of measurements. What shape does this distribution resemble? Let us assume that this distribution is Gaussian with mean $\mu(d)$ for the actual distance d and variance $\sigma_s^2(d)$ which could itself depend on the distance. The noise model can then be written as

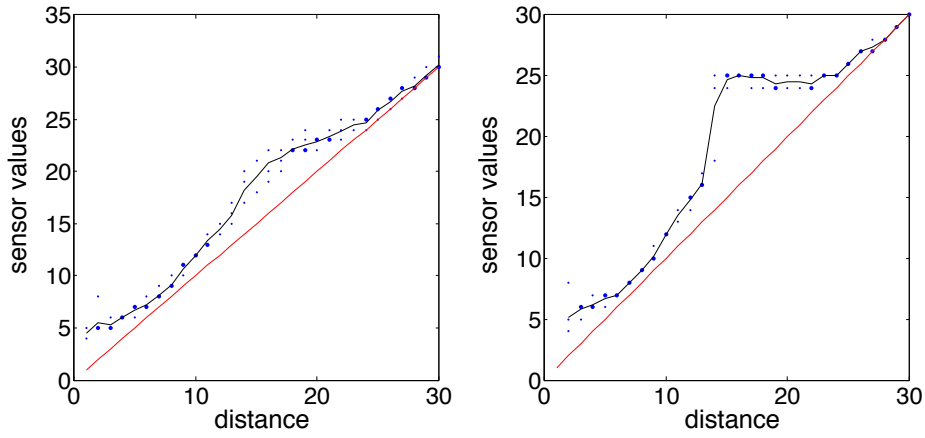


Fig. 3.3 Example of response a sensor function for two different ultrasonic sensors from a test setup as shown in Fig. 3.4. The dots show the different readings in repeated trials, and their size indicates the frequency with which this particular value is encountered. Some of the tests included a systematic variation of the true distance, whereas others chose a random order of different distances.

$$p(x|Z;\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{x^2}{2\sigma_s^2}}, \tag{3.40}$$

Where $x(d)$ is a function that we need to characters further in order to make some predictions that we can take into account when using the sensor data. Finding this function, or more precisely the density function that describes the training examples, is the major focus of supervised learning that we discuss in the next section. For now, we should note that the measurement model encapsulates all the information we can get from a sensor.

3.3.1 Probabilistic motion models

Sensors are not the only noisy parts of a robot. Indeed, the motion of a robot is commonly even more unreliable. We thus need a motion model that takes uncertainties into account and that returns the probability of a new pose x_t after applying a motor command m at $t-1$. The new state might depend on the history of previous states, $\{x_0, ..., x_{t-1}\}$, but we make here the common **Markov assumption** that the new state only depends on the previous state,

$$\textbf{Motion model: } p(x_t|x_{t-1}, m). \tag{3.41}$$

We already discussed the deterministic part of dynamic movement in the kinematic models of the last chapter, and we will here augment these within a probabilistic framework. To illustrate this, let's use a simple sample of a tribot moving on a one-dimensional trajectory. The position on this trajectory is specified by the position variable x. We want to apply a specific motor commands that will turn on both motors

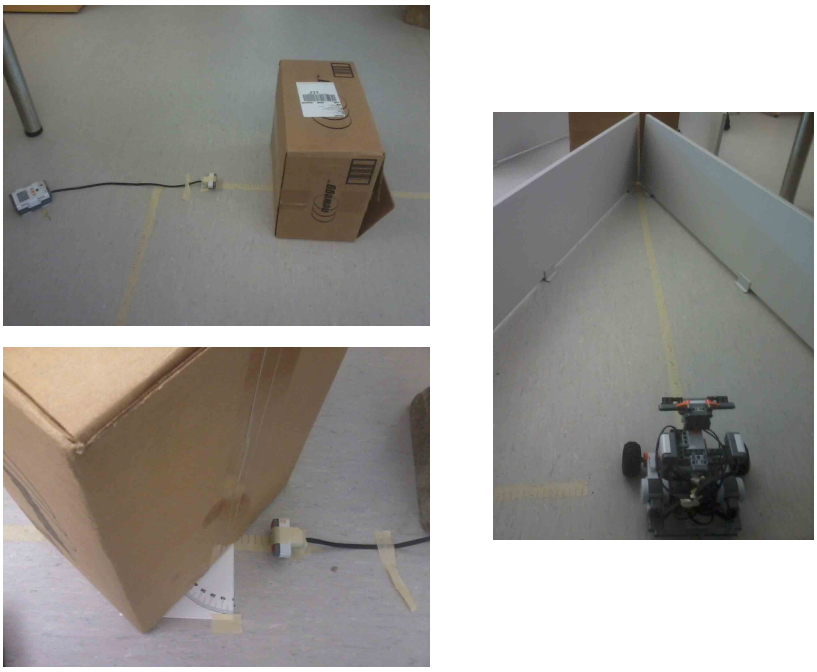


Fig. 3.4 Setup of experiments to evaluate the ultrasonic sensor. Note that the performance of the ultrasonic sensor in a corner situation is quite problematic.

for a specific time t_m. The displacement of the robot is then to a first approximation linear in this time,

$$\Delta x = x_t - x_{t-1} = a_0 + a\, t_m. \tag{3.42}$$

This is a linear kinematic model where we included a constant a_0 to describe the effect of a latency when applying the motor command. If we only take this kinematic model into account to calculate the new position of the robot without any sensor feedback, then this is often called **dead reckoning** in navigation.

While the kinematic model give us a baseline for what to position to expect after applying a motor command, we also know that movements will introduce errors such as from slippage of the wheels or external factors that alters the position of the robot. Indeed, taking uncertainties into account, such as noisy movements, unexpected environments, or inconsistent orders, is a crucial step to make robots work in the real world.

To capture at least some kind of uncertainty, we can run the tribot repeatedly for a specific time and measure the distance traveled. One can then plot a histogram of these positions to estimate the noise distribution. Let us assume again that this noise is Gaussian. The motion model can then be written as

$$p(x_t|x_{t-1}, t_m; \theta) = x_{t-1} + a_0 + at_m + \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(b_m)^2}{2\sigma_m^2}}, \tag{3.43}$$

with parameters b_m and σ_m that describe the Gaussian noise that we used here to describe noisy observation. Of course, we need to measure examples of real movements

to see if they are Gaussian or if other noise models would be more appropriate. The next chapters will introduce techniques to learn such models from examples.

Exercise

1. Use the light sensor to measure distances to a surface and derive a sensor model for this sensor. Provide a parametric form of your model and include estimations of the parameters.
2. Derive a motion model for the tribot when driving the motors with different power parameters. Provide a parametric form of your model and include estimations of the parameters.