Chapter 2
Probability Theory and Random Variables

This chapter collects the basic terms from probability theory and statistics. It motivates the axiomatic approach for the concept of probability, introduces the concept of a random variable, describes the key properties of the main distributions of random variables occurring when modelling observational uncertainties and testing hypotheses, and provides an introduction to stochastic processes. We give the key methods for determining the uncertainty of derived entities, especially for explicit and implicit functions of single and multiple variables. The reader who has had a basic course on statistics may take a quick look at the notation used and the lines of thought employed. The concepts can be found in the excellent textbooks by Papoulis (1965) and Papoulis and Pillai (2002) and online at `http://www.math.uah.edu/stat/index.html`.

2.1 Notions of Probability

Probability theory is the most powerful tool for working with uncertainty. The notion of probability has changed over the last two centuries.

- The *classical definition* of probability P according to Laplace is the ratio of the number n_+ of favourable to the number n of possible cases of an event $\mathcal{E}$,

$$P(\mathcal{E}) \doteq \frac{n_+}{n}\,. \tag{2.1}$$

  When modelling the outcome of throwing a die, e.g., this definition leads to the usually assumed probability $1/6$ for each possible event.
  But when modelling the outcome of a modified die, e.g., one that yields more sixes, we encounter difficulties with this definition. We would need to define conditions for the different events under which they occur with the same probability, thus requiring the notion of probability.
  In the case of alternatives which are not countable, e.g., when the event is to be represented by a real number, we have difficulties in defining equally probable events.

© Springer International Publishing Switzerland 2016

W. Förstner and B.P. Wrobel, *Photogrammetric Computer Vision*, Geometry and Computing 11,
DOI 10.1007/978-3-319-11550-4_2

Bertrand's paradox

This is impressively demonstrated by *Bertrand's paradox* (Fig. 2.1), which answers the question: What is the probability of an *arbitrarily chosen* secant in a circle longer than the side of an inscribing equilateral triangle? We have three alternatives for specifying the experiment:

1. Choose an arbitrary point in the circle. If it lies within the concentric circle with half the radius, then the secant having this point as centre point is longer than the sides of the inscribing triangle. The probability is then $1/4$.
2. Choose an arbitrary point on the circle. The second point of the secant lies on one of the three segments inducing sectors of $60°$. If the second point lies in the middle sector the secant through these points is longer than the side of the inscribing triangle. The probability is then $1/3$.
3. Choose an arbitrary direction for the secant. If its centre point lies in one of the two centre quarters of the diameter perpendicular to this direction the secant is longer than the side of the inscribing triangle. The probability is then $1/2$.

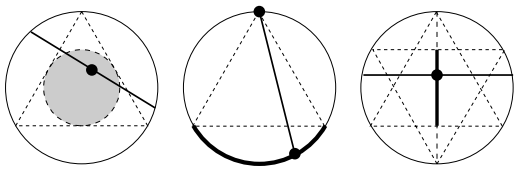

Fig. 2.1 Bertrand's paradox: Three alternatives for choosing an arbitrary secant in a circle. **Left:** choosing an arbitrary point in the small circle with half radius, and interpreting it as the middle of the secant; **Middle:** by first choosing a point on the boundary, then the second point must lie in a certain range of the boundary, namely in between the secants belonging to an equilateral triangle; **Right:** choosing an arbitrary point on a diameter, in the middle range of the secant

Obviously the definition of the notion *arbitrarily chosen*, i.e., an equal probability, is not simple. However, this definition is often used, as it follows the classical logic under certain conditions.

- The definition of probability as *relative frequency* following von Mises. This definition follows the empirical finding that the empirical relative frequency seems to converge to a limiting value
$$P(\mathcal{E}) \doteq \lim_{n\to\infty} \frac{n_+}{n}\,. \tag{2.2}$$
This plausible definition fails in practice, as the number of experiments will not be sufficiently large and the conditions for an experiment cannot be held stable over a long enough time.
- Probability as the degree of *subjective certainty*, e.g., in the sentence: *"There is a large probability this statement, A, is correct."*
Due to its subjectivity, this definition is not suitable as a basis for a theory. However, sometimes we use subjective probabilities, which then requires a rigorous definition of the concept.

All three definitions are plausible and form the basis for the following axiomatic definition.

2.2 Axiomatic Definition of Probability

The following axiomatic definition of probability follows Kolmogorov and solves the issues of the previous definitions (Fig. 2.2).

Kolmogorov's Axiomatic Definition of Probability. Basis is a space S of elementary events $A_i \in S$. Events A are subsets of S. The certain event is S, the impossible event is $\emptyset$. Each combination of events A and B again is an event; thus, the alternative event $A \cup B$, the joint event $A \cap B$ and the negated event $\overline{A} = S - A$ are events.

Each event can be characterized by a corresponding number, $P(A)$, its probability, which fulfils the following three axioms:

axiomatic definition of probability

1. For any event, we have
$$P(A) \geq 0\,. \tag{2.3}$$
2. The certain event has probability 1,
$$P(S) = 1\,. \tag{2.4}$$
3. For two mutually exclusive events, $A \cap B = \emptyset$ (Fig. 2.2, a),
$$P(A \cup B) = P(A) + P(B)\,. \tag{2.5}$$

Conditional Probability. Moreover, we have the conditional probability of an event A given the event B has occurred. The probability
$$P(A \mid B) = \frac{P(A,B)}{P(B)} \tag{2.6}$$
is the ratio of the joint probability $P(A,B) = P(A \cap B)$ of events A and B occurring simultaneously and the probability $P(B)$ of only B occurring (Fig. 2.2, b).

Total Probability. The total probability of an event A in the presence of a second event $B = \bigcup_{i=1}^{I} B_i$ therefore is (Fig. 2.2, c)
$$P(A) = \sum_{i=1}^{I} P(A \mid B_i) P(B_i)\,. \tag{2.7}$$

Independent Events. Two events A and B are called independent (Fig. 2.2, d) if
$$P(A,B) = P(A)P(B)\,. \tag{2.8}$$

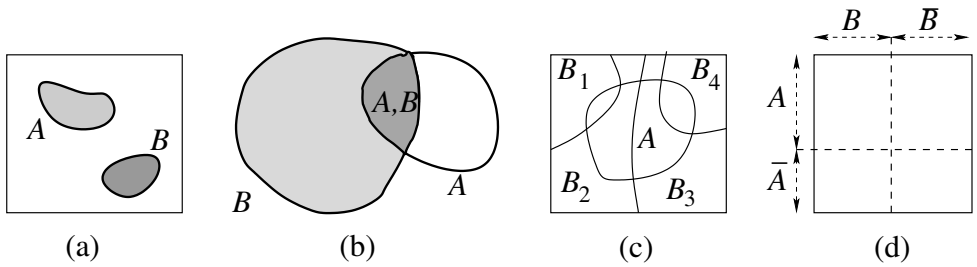

Fig. 2.2 Independence, conditional and total probability. (a) Disjoint events A and B, (b) conditional probability $P(A \mid B) = P(A,B)/P(B)$, (c) total probability $P(A)$, (d) independent events A and B

These axioms coincide with the classical definition of probability if the definition of elementary events is unique and can be considered as equally probable.

Example 2.2.1: ***Throwing a die.*** (1) When throwing a die, we have the set S of elementary events

$$S = \{s_1, s_2, s_3, s_4, s_5, s_6\},$$

and for each s_i

$$P(s_i) = \frac{1}{6}.$$

(2) Throwing two dice i, j, we have

$$S = \{(s_i, s_j)\} \quad \text{and} \quad P((s_i, s_j)) = \frac{1}{36}.$$

(3) The *conditional probability* $P(s_2 \mid \mathsf{even})$ of throwing a 2, i.e., event s_2, under the condition that we know that an even number was thrown, and using (2.6) is

$$P(s_2 \mid \{s_2, s_4, s_6\}) = \frac{P(s_2, \{s_2, s_4, s_6\})}{P(\{s_2, s_4, s_6\})} = \frac{P(s_2)}{P(\{s_2, s_4, s_6\})} = \frac{1/6}{1/2} = \frac{1}{3}.$$

(4) The *total probability* for $A = \mathsf{even} := \{s_2, s_4, s_6\}$ (i.e., throwing an even number) if having thrown a $B_1 = \mathsf{small}$ or $B_2 = \mathsf{large}$ number (with $B = \{\mathsf{small}, \mathsf{large}\} := \{\{s_1, s_2, s_3\}, \{s_4, s_5, s_6\}\}$), is

$$P(\mathsf{even}) = P(\mathsf{even} \mid \{s_1, s_2, s_3\})P(\{s_1, s_2, s_3\}) + P(\mathsf{even} \mid \{s_4, s_5, s_6\})P(\{s_4, s_5, s_6\}) = \frac{1}{3}\frac{1}{2} + \frac{2}{3}\frac{1}{2} = \frac{1}{2}.$$

(5) The events $A = s_1$ to first throw 1 and $B = \{s_2, s_4, s_6\}$ to secondly throw even are *independent*;

$$\begin{aligned} P(1, \mathsf{even}) &= P(\{(s_1, s_2), (s_1, s_4), (s_1, s_6)\}) = \frac{3}{36} \\ &= P(1)P(\mathsf{even}) = \frac{1}{6}\frac{1}{2} = \frac{1}{12}. \end{aligned}$$

◇

2.3 Random Variables

random variables for unifying numerical and nonnumerical experimental outcomes

For experiments with a nonnumerical outcome, e.g., a colour, it is useful to map the experimental outcome to a real value and describe the probabilistic properties of the experiment using a real-valued *random variable*.

Since such a mapping in a natural way can be defined for experiments with discrete or continuous outcome, random variables in a unifying manner play a central role in stochastic modelling.

2.3.1 Characterizing Probability Distributions

With each outcome $s \in S$ of an experiment, we associate a real number $\underline{x}(s) \in \mathbb{R}$. The function $\underline{x}$

$$\underline{x} : S \to \mathbb{R} \qquad \underline{x} = \underline{x}(s) \tag{2.9}$$

is called a random variable. In order to specify the randomness of the experiment, thus, instead of characterizing the possible outcomes s, we characterize the function $\underline{x}$ (cf. Papoulis and Pillai, 2002). Observe: we distinguish between a sample value $x(s)$ (without underscore) depending on the outcome s of a *specific* experiment and the random variable $\underline{x}(s)$ (with underscore) which describes the experiment *as a whole*, for all $s \in S$. We regularly denote the random variable by $\underline{x}$, omitting the dependency of s.

Specifically, the experiment is characterized by what is called the *distribution* or *probability function*,

$$P_x(x) = P(\underline{x} < x)\,. \tag{2.10}$$

The argument $\underline{x} < x$ is the set of all possible outcomes for which $\underline{x}(s) < x$ holds. This definition assumes that there exists an event for all $x \in \mathbb{R}$.

The *index* x in the probability function $P_x(x)$ refers to the associated random variable, whereas the *argument* in $P_x(x)$ is the variable of the function. For simplicity, we sometimes omit the index.

observation process characterized by random variables

We will regularly characterize the statistical properties of an *observation process* by one or more random variables, catching that aspect of the concrete observation procedure which is relevant for the analysis task.

We can now derive the probability of a random variable to be in an interval,

$$P(\underline{x} \in [a, b]) = P_x(b) - P_x(a)\,. \tag{2.11}$$

Obviously, a probability function must fulfil

- $P_x(-\infty) = 0$,
- $P_x(x)$ is not decreasing, and
- $P_x(\infty) = 1$.

Example 2.3.2: ***Throwing a coin.*** When throwing a coin, we assume that

$$\underline{x}(\mathsf{heads}) = 0 \qquad \underline{x}(\mathsf{tails}) = 1\,. \tag{2.12}$$

In the case of equal probability of each outcome, we obtain the probability function

$$P_c(x) = \begin{cases} 0 & \text{if } x \le 0 \\ 1/2 & \text{if } 0 < x \le 1 \\ 1 & \text{else} \end{cases}\,. \tag{2.13}$$

Observe, the index c in P_c is part of the name P_c of the probability function, here referring to throwing a coin. For the range $x \in (-\infty, 0]$, the corresponding event is the empty set $\emptyset$: it is unlikely that throwing a coin leads to neither heads nor tails. For the range $x \in (0, 1]$, the corresponding event is heads as $P(\underline{x}(\mathsf{heads}) < x) = 1/2$. For the range $x \in (1, \infty)$, the corresponding event is the certain event S. The probability of the event tails is given by $P(\mathsf{tails}) = P(\neg\mathsf{heads}) = 1 - P(\mathsf{heads}) = 1/2$, as the events heads and tails are mutually exclusive. Thus the event tails cannot be represented by some interval. ◇

unit step function

Using the unit-step function $s(x)$ (Fig. 2.3),

$$s(x) = \begin{cases} 0 & \text{if } x \le 0, \\ 1 & \text{else} \end{cases} \tag{2.14}$$

the probability function P_c can be written as

$$P_c(x) = \frac{1}{2}s(x) + \frac{1}{2}s(x-1)\,. \tag{2.15}$$

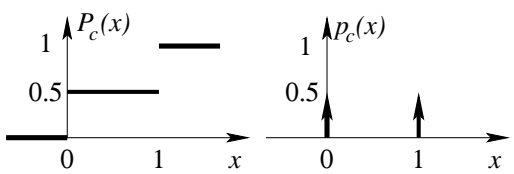

Fig. 2.3 Probability function $P_c(x)$ and density function $p_c(x)$ for throwing a coin

2.3.2 Probability Density Function

For experiments with continuous outcomes, e.g., a length measurement, we usually choose[1] $\underline{x}(x) = x$. We characterize the experiment by the first derivative of the probability function, which is called the *probability density function* or just *density function*

probability density function or density function

$$p_x(x) = \frac{\mathrm{d}P_x(x)}{\mathrm{d}x}\,. \tag{2.16}$$

Since integrating $p_x(x)$ yields $P_x(x)$ (cf. (2.10), p. 25)

$$P_x(x) = \int_{t=-\infty}^{x} p_x(t)\,\mathrm{d}t\,. \tag{2.17}$$

The function $P_x(x)$ is also is called the *cumulative distribution function* or just *cumulative distribution*. It is the same function as in (2.10), p. 25.

Example 2.3.3: ***Rounding errors.*** Rounding errors e lie in the interval $[-\frac{1}{2}, \frac{1}{2}]$. The probability of a rounding error to lie in the subinterval $[a, b] \subset [-\frac{1}{2}, \frac{1}{2}]$ is proportional to the ratio of the length $b - a$ to the length 1 of the complete interval. Therefore the probability density is

$$p_e(x) = r\left(x\middle| -\frac{1}{2}, \frac{1}{2}\right) \doteq \begin{cases} 1 & \text{if } x \in \left[-\frac{1}{2}, \frac{1}{2}\right] \\ 0 & \text{otherwise.} \end{cases} \tag{2.18}$$

This is the density of the uniform distribution in the interval $[-\frac{1}{2}, \frac{1}{2}]$, see Fig. 2.4. ◇

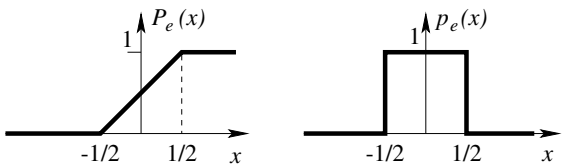

Fig. 2.4 Probability distribution $P_e(x)$ and probability density function $p_e(x)$ of the rounding error $\underline{e}$

2.3.3 Continuous and Discrete Random Variables

Random variables are called *continuous* if their probability distribution is continuous or, equivalently, if their density function is bounded. A random variable is called *discrete* if the probability function contains only steps or, equivalently, if the probability density function is either zero or infinite at a countable number of values x.

Example 2.3.4: ***Discrete probability density function.*** The probability density function of the random variable $\underline{x}$ of tossing a coin is

$$p_x(x) = \frac{1}{2}\delta(x) + \frac{1}{2}\delta(x-1)\,,$$

where $\delta(x)$ is Dirac's delta function. ◇

Dirac's delta function is the first derivative of the unit step function

$$\delta(x) \doteq \frac{\mathrm{d}s(x)}{\mathrm{d}x} \tag{2.19}$$

and is defined by a limiting process, e.g., by:

[1] The random variable depends on the unit in which x is measured, e.g., m or cm.

$$\delta(x) = \lim_{d\to 0} r(x|-d,+d) \tag{2.20}$$

with the rectangle function

$$r(x|a,b) \doteq \begin{cases} \dfrac{1}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{else} \end{cases} . \tag{2.21}$$

The Dirac function has the following properties: The area under the delta function is 1:

$$\int_{t=-\infty}^{\infty} \delta(t)\mathrm{d}t = \lim_{x\to 0} \int_{t=-x}^{x} \delta(t)\mathrm{d}t = \lim_{x\to 0}(s(x)-s(-x)) = 1\,. \tag{2.22}$$

Therefore,

$$\int_{t=-\infty}^{\infty} f(x-t)\delta(t)\mathrm{d}t \overset{t\to \underline{x}-t}{=} \int_{t=-\infty}^{\infty} f(t)\delta(x-t)\mathrm{d}t \tag{2.23}$$

$$\overset{\delta(x)=0 \text{ for } x\neq 0}{=} \lim_{d\to 0} \int_{t=x-d}^{x+d} f(t) r(x|t-d,t+d)\mathrm{d}t \tag{2.24}$$

$$\overset{\xi\in[t-d,t+d]}{=} \lim_{d\to 0} \int_{t=x-d}^{x+d} \frac{1}{2d} f(\xi)\mathrm{d}t = f(x), \tag{2.25}$$

the second last step using the mean value theorem for integration. The delta function can thus be used to select a certain value of a function $f(x)$.

In graphs, the delta function is visualized as an *arrow* with the height indicating the local area under the function. For discrete random variables, therefore, we draw the heights of these *arrows*, i.e., the probabilities that one of the countable number of events occurs. Instead of the density function $p_x(x) = 1/2\,\delta(x) + 1/2\,\delta(x-1)$ for tossing a coin, e.g., we give the two probabilities $P(x=0) = P(x=1) = 1/2$.

The distribution of a random variable is often given a name, e.g., $\mathcal{H}$, and we write $\underline{x} \sim \mathcal{H}$ or, if the distribution depends on parameters $\boldsymbol{p}$,

$$\underline{x} \sim \mathcal{H}(\boldsymbol{p})\,. \tag{2.26}$$

2.3.4 Vectors and Matrices of Random Variables

We often have experiments with multiple outcomes. The corresponding I random variables $\underline{x}_i$ are usually collected in a vector called a *random vector*,

$$\underline{\boldsymbol{x}} = [\underline{x}_1, ..., \underline{x}_i, ..., \underline{x}_I]^\mathsf{T}\,. \tag{2.27}$$

The experiment is then characterized by the multi-dimensional probability function

$$P_{x_1,...,x_i,...,x_I}(\underline{x}_1 \leq x_1, ..., \underline{x}_i \leq x_i, ..., \underline{x}_I) = P(x_1, ..., x_i, ..., x_I) \tag{2.28}$$

or

$$P_x(\underline{\boldsymbol{x}} \leq \boldsymbol{x}) = P(\boldsymbol{x}), \tag{2.29}$$

or by the multi-dimensional probability density function

$$p_x(\boldsymbol{x}) = \frac{\partial^I P(\boldsymbol{x})}{\partial x_1 ... \partial x_i ... \partial x_I}\,. \tag{2.30}$$

random matrices

We will regularly use random matrices, e.g., when dealing with uncertain transformations. Let the $N \times M$ matrix $\underline{\boldsymbol{X}} = [X_{nm}]$ contain NM random variables. Then it is of

advantage to vectorize the matrix,

$$\underset{NM\times 1}{\underline{\boldsymbol{x}}} = \mathrm{vec}\underline{X} = [\underline{X}_{11}, \underline{X}_{21}, ..., \underline{X}_{N1}, \underline{X}_{12}, ..., \underline{X}_{NM}]^{\mathsf{T}} \tag{2.31}$$

and represent the uncertainty by the joint probability of the random NM-vector $\underline{\boldsymbol{x}}$.

2.3.5 Statistical Independence

If two random variables $\underline{x}$ and $\underline{y}$ are statistically independent, their joint probability function and their joint probability density function are separable functions, i.e.,

$$P_{xy}(x, y) = P_x(x)P_y(y) \quad \text{or} \quad p_{xy}(x, y) = p_x(x)p_y(y)\,. \tag{2.32}$$

2.4 Distributions

We now list a number of distributions relevant for statistical reasoning.

2.4.1 Binomial Distribution

A discrete random variable $\underline{n}$ follows a binomial distribution,

$$\underline{n} \sim \mathrm{Bin}(N, p), \tag{2.33}$$

if its discrete density function is

$$P(n) = \binom{N}{n} p^n (1-p)^{N-n} \qquad n = 0, 1, ..., N \qquad 0 \leq p \leq 1 \tag{2.34}$$

where $\binom{N}{n}$ are binomial coefficients. It models the probability of n successes if an experiment for which the probability of success p is repeated N times.

For $p = \frac{1}{2}$, we obtain the probability $P(n)$ of observing n heads when tossing a coin N times (Table 2.1).

2.4.2 Uniform Distribution

A continuous random variable follows the general uniform distribution,

$$\underline{x} \sim \mathcal{U}(a, b) \qquad a, b \in \mathbb{R} \quad b > a, \tag{2.35}$$

Table 2.1 Probability $P(n)$ of obtaining n heads when tossing a coin N times

	$n=0$	1	2	3	4	5	6
$N=1$	$\frac{1}{2}$	$\frac{1}{2}$					
2	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$				
3	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$			
4	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{16}$		
5	$\frac{1}{32}$	$\frac{5}{32}$	$\frac{5}{16}$	$\frac{5}{16}$	$\frac{5}{32}$	$\frac{1}{32}$	
6	$\frac{1}{64}$	$\frac{3}{32}$	$\frac{15}{64}$	$\frac{5}{16}$	$\frac{15}{64}$	$\frac{3}{32}$	$\frac{1}{64}$

if it has the density $r(x|a,b)$ ((2.21), p. 27). For example, *rounding errors* $\underline{e}$ have uniform distribution $\underline{e} \sim \mathcal{U}(-\frac{1}{2}, \frac{1}{2})$.

Two random variables $\underline{x}$ and $\underline{y}$ jointly follow a uniform distribution,

$$(\underline{x}, \underline{y}) \sim \mathcal{U}(a, b; c, d), \tag{2.36}$$

if they have the density function

$$r_{xy}(x, y \mid a, b; c, d) = r(x \mid a, b)\; r(y \mid c, d)\,, \tag{2.37}$$

where $x \in [a, b]$ and $y \in [c, d]$. Due to (2.37) the random variables $\underline{x}$ and $\underline{y}$ are independent.

2.4.3 Exponential and Laplace Distribution

A random variable $\underline{x}$ follows an exponential distribution with real parameter $\mu > 0$ if its density function is given by

$$p_x(x) = \frac{1}{\mu} \mathrm{e}^{-\dfrac{x}{\mu}}\,, \qquad x \geq 0, \quad \mu > 0\,. \tag{2.38}$$

This is also called the *Rayleigh distribution*. *Rayleigh distribution*

A random variable $\underline{x}$ is Laplacian distributed with real parameter $\sigma > 0$,

$$\underline{x} \sim \mathrm{Lapl}(\sigma), \tag{2.39}$$

if its density function is given by

$$p_x(x) = \frac{1}{\sqrt{2}\,\sigma} \mathrm{e}^{-\sqrt{2}\left|\frac{x}{\sigma}\right|}\,, \qquad \sigma > 0\,. \tag{2.40}$$

2.4.4 Normal Distribution

2.4.4.1 Univariate Normal distribution

A random variable $\underline{x}$ is normally or Gaussian distributed with real parameters μ and $\sigma > 0$,

$$\underline{x} \sim \mathcal{N}(\mu, \sigma^2), \tag{2.41}$$

if its density function is given by

$$p_x(x) = g(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \qquad \sigma > 0\,. \tag{2.42}$$

The density function is symmetric with respect to μ, there having the value $1/(\sqrt{2\pi}\,\sigma) \approx 0.4/\sigma$; the inflection points are at $\mu - \sigma$ and $\mu + \sigma$, there having the value $1/(\sqrt{2\pi\mathrm{e}}\,\sigma) \approx 0.24/\sigma$, hence 3/5th of the value at the mean. The tangents at the inflection points intersect the x-axis at $\mu \pm 2\sigma$.

Large deviations from the mean value μ are unlikely:

$$P(\underline{x} \in [\mu - \sigma, \mu + \sigma]) = \int_{x=\mu-\sigma}^{x=\mu+\sigma} g(x \mid \mu, \sigma^2)\,\mathrm{d}x \approx 0.6827\,, \tag{2.43}$$

$$P(\underline{x} \in [\mu - 2\sigma, \mu + 2\sigma]) = \int_{x=\mu-2\sigma}^{x=\mu+2\sigma} g(x \mid \mu, \sigma^2)\,\mathrm{d}x \approx 0.9545\,, \tag{2.44}$$

$$P(\underline{x} \in [\mu - 3\sigma, \mu + 3\sigma]) = \int_{x=\mu-3\sigma}^{x=\mu+3\sigma} g(x \mid \mu, \sigma^2)\,\mathrm{d}x \approx 0.9973\,. \tag{2.45}$$

Thus the probability of a value lying outside the interval $[\mu-3\sigma, \mu+3\sigma]$ is very low, 0.3 %.

standard normal distribution, normalized Gaussian distribution

The *standard normal distribution* or *normalized Gaussian distribution* is given by $\mu = 0$ and $\sigma = 1$ (Fig. 2.5)

$$\phi(x) = g(x \mid 0, 1) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}\,. \tag{2.46}$$

Its cumulative distribution is

$$\Phi(x) = \int_{t=-\infty}^{x} \phi(t)\,\mathrm{d}t\,. \tag{2.47}$$

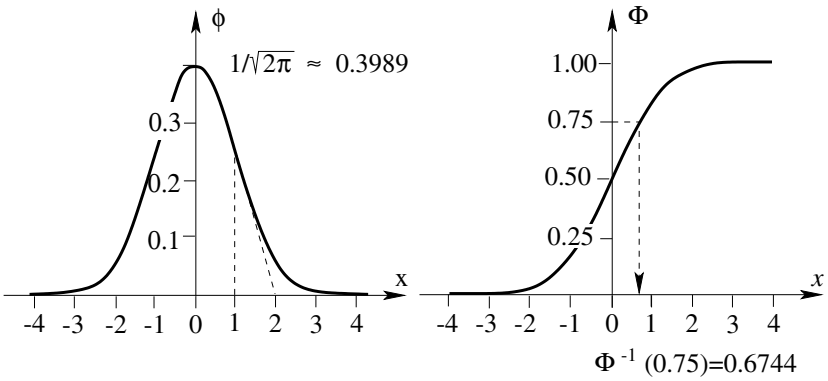

Fig. 2.5 Left: normal or Gaussian density function $\phi(x)$. Inflection points at $x = +1$ and $x = -1$. The ratio of the function values on the symmetry axis and at the inflection point is $\sqrt{\mathrm{e}} = 1.6487... \approx 5/3$; the tangent in the inflection point intersects the x-axis at $x = 2$, such that the x-coordinate of the inflection point is in the middle of this intersection point and the line of symmetry. **Right:** cumulative distribution function $\Phi(x)$. 75th percentile at $x = \Phi^{-1}(0.75) = 0.6745$

central limit theorem

The normal distribution is the most important distribution. This follows from the *central limit theorem*: The sum of a large number of independent, identically distributed random variables with bounded variance is approximately normally distributed (cf. Papoulis, 1965, Sect. 8–6).

2.4.4.2 Multi-dimensional Normal Distribution

If two independent random variables are normally distributed according to

$$\underline{x} \sim \mathcal{N}(\mu_x, \sigma_x^2) \qquad \underline{y} \sim \mathcal{N}(\mu_y, \sigma_y^2), \tag{2.48}$$

their joint density function is

$$p_{xy}(x,y) = g_x(x \mid \mu_x, \sigma_x^2)\; g_y(y \mid \mu_y, \sigma_y^2) \tag{2.49}$$

$$= \frac{1}{2\pi\sigma_x\sigma_y} \mathrm{e}^{-\frac{1}{2}\left(\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right)}. \tag{2.50}$$

With the vectors

$$\boldsymbol{x} = \begin{bmatrix} x \\ y \end{bmatrix} \qquad \mu = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} \tag{2.51}$$

and the 2×2 matrix,

$$\Sigma = \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix}, \tag{2.52}$$

this can be written as

$$g_{xy}(\boldsymbol{x} \mid \boldsymbol{\mu}, \Sigma) = \frac{1}{2\pi\sqrt{|\Sigma|}} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\mathsf{T}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}. \tag{2.53}$$

If the 2×2 matrix Σ is a general symmetric positive definite matrix

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix}, \tag{2.54}$$

the two random variables are dependent. The correlation coefficient,

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x\sigma_y} \in [-1, 1], \tag{2.55}$$

measures the degree of linear dependency. If $\rho_{xy} = 0$, the two random variables are *uncorrelated*, and if they are normally distributed, they are *independent*, due to (2.32), p. 28. The 2D normal distribution is an elliptic bell-shaped function and can be visualized by one of its contour lines, cf. Fig. 2.6. The *standard ellipse*, sometimes also called standard error ellipse, is defined by

uncorrelated, independent random variables

standard ellipse

$$(\boldsymbol{x}-\boldsymbol{\mu})^\mathsf{T}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = 1\,. \tag{2.56}$$

The standard ellipse allows the visualization of important properties of the uncertain point:

- The standard ellipse is centred at $\boldsymbol{\mu}_x$.
- The bounding box has size $2\sigma_x \times 2\sigma_y$.
- The semi-axes are the square roots of the eigenvalues λ_i of the covariance matrix, namely $\sigma_{\max} = \sqrt{\lambda_1}$ and $\sigma_{\min} = \sqrt{\lambda_2}$, which are the square roots of the eigenvalues of Σ,

$$\sigma^2_{\max,\min} = \frac{1}{2}(\sigma_x^2 + \sigma_y^2) \pm \frac{1}{2}\sqrt{(\sigma_x^2 - \sigma_y^2)^2 + 4\sigma_{xy}^2}\,. \tag{2.57}$$

- If the two coordinates are correlated, the major axis is not parallel to the coordinate system. The angle α is given by

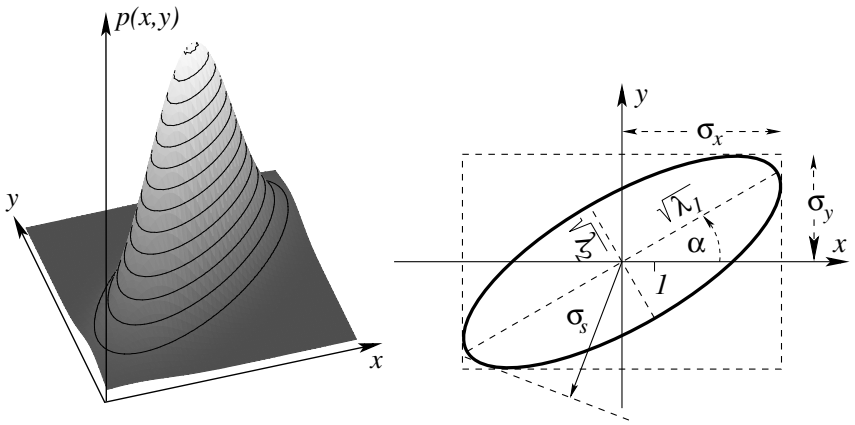

Fig. 2.6 General 2D normal or Gaussian distribution, centred at the origin. **Left:** density function. **Right:** standard ellipse. Actual values: $\mu_x = \mu_y = 0$, $\sigma_x = 4.9$, $\sigma_y = 3.2$, $\rho = 0.7$

$$\alpha = \frac{1}{2}\text{atan2}\,(2\sigma_{xy}, \sigma_x^2 - \sigma_y^2) \in (-\pi/2, +\pi/2] \tag{2.58}$$

using a two-argument version of the arctan function.

The sign of the angle follows the sign of the correlation coefficient ρ_{xy} or the covariance σ_{xy}.

- The standard deviation σ_s of a distance s between the point $\boldsymbol{\mu}_x$ and a fixed point in an arbitrary direction, indicated here by an arrow, is given by the distance of $\boldsymbol{\mu}_x$ from the tangent to the standard ellipse perpendicular to that direction. This shows that the minor and the major axes of the standard ellipse give the minimum and the maximum of the directional uncertainty of the point.

In higher dimensions, (2.56) represents an ellipsoid or a hyper-ellipsoid $\mathcal{E}$. The probability $S = P(\underline{x} \in \mathcal{E})$ that a random point lies within the standard ellipsoid depends on the dimension as shown in the first line of Table 2.2, and rapidly diminishes with the dimension.

confidence ellipse

Instead of showing the standard ellipse or standard ellipsoid, we therefore can show the *confidence ellipse* or *confidence ellipsoid*. The confidence ellipsoid is the k-fold standard ellipsoid, such that the probability $P(\underline{x} \in \mathcal{E}(k))$ that a sample lies within the ellipsoid is a certain prespecified value S

$$\mathcal{E}(k): \quad (\boldsymbol{x} - \boldsymbol{\mu})^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = k^2\,, \quad P(\underline{x} \in \mathcal{E}(k)) = S\,. \tag{2.59}$$

The standard ellipse is identical to the confidence ellipse for $k = 1$: $\mathcal{E} = \mathcal{E}(1)$. For the dimension $d = 1$ and a probability $P(\underline{x} \in \mathcal{E}(k)) = S = 0.9973$, we would obtain $k = 3$, as shown in (2.45), p. 30. Here the ellipse reduces to the interval $[-k\sigma_x, +k\sigma_x]$.

For $S = 95\%$, $S = 99\%$ and $S = 99.9\%$, the values $k(S)$ determined from the right equation in (2.59) are given in Table 2.2 for different dimensions.

Table 2.2 Confidence regions. **First row:** Probabilities $P(\underline{x} \in \mathcal{E})$ for different dimensions d of a random vector $\underline{x}$. **Other rows:** Factor $k(S)$ for the confidence ellipsoids $\mathcal{E}(k(S))$ for $S = 0.95, 0.99, 0.999$ and for different dimensions d.

d	1	2	3	4	5	10	20	50	100
$P(\underline{x} \in \mathcal{E})$	0.68	0.40	0.20	0.09	$3.7 \cdot 10^{-2}$	$1.7 \cdot 10^{-4}$	$1.7 \cdot 10^{-10}$	$1.2 \cdot 10^{-33}$	$1.8 \cdot 10^{-80}$
$k(0.95)$	1.96	2.45	2.80	3.08	3.33	4.28	5.60	8.22	11.2
$k(0.99)$	2.58	3.03	3.37	3.64	3.88	4.82	6.13	8.73	11.6
$k(0.999)$	3.29	3.72	4.03	4.30	4.53	5.44	6.73	9.31	12.2

Gaussian distributed matrix

Matrices of Gaussian distributed random variables can be represented using their vector representation, (2.31), p. 28. Let the $N \times M$ matrix $\underline{\boldsymbol{X}}$ contain NM random variables which are normally distributed; we represent its uncertain covariance matrix using the

random vector

$$\underline{\boldsymbol{x}} = \mathrm{vec}\underline{X}: \quad \underline{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{\mu}_x, \Sigma_{xx})\,. \tag{2.60}$$

Or we may keep the matrix representation for the mean matrix and write

$$\underline{X} \sim \mathcal{N}(\boldsymbol{\mu}_X, \Sigma_{xx})\,. \tag{2.61}$$

Sometimes we will refer to Σ_{xx} as the *covariance matrix of the random matrix* $\underline{X}$.

2.4.4.3 Normal Distribution with Zero or Infinite Variance

When representing fixed values, such as the third component in a homogeneous vector $[\underline{x}, \underline{y}, 1]^\mathsf{T}$, we might track this property through the reasoning chain, which is cumbersome, or just treat the value 1 as a stochastic variable with mean 1 and variance 0. The second alternative has implicitly been chosen by Kanatani (1996) and Criminisi (2001). This method needs some care, as the density function for a Gaussian random variable is not defined for zero variance.

The distribution of a random variable $\underline{y} \sim \mathcal{N}(\mu_y, 0)$ can be defined in a limiting process ((2.22), p. 27), by a δ-function:

$$p_y(y) = \lim_{\sigma_y \to 0} g(y; \mu_y, \sigma_y^2) = \delta(y - \mu_y)\,. \tag{2.62}$$

Now a 2-vector can be constructed with a singular 2×2 covariance matrix. Assume that $\underline{x} \sim N(\mu_x, 1)$ and $\underline{y} \sim N(\mu_y, 0)$ are independent stochastic variables; thus,

$$\begin{bmatrix} \underline{x} \\ \underline{y} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \right)\,. \tag{2.63}$$

As $\underline{x}$ and $\underline{y}$ are stochastically independent, their *joint generalized probability density function* is ((2.32), p. 28)

$$g_{xy} = g_x(x; \mu_x, 1)\ \delta(y - \mu_y)\,. \tag{2.64}$$

Obviously, working with a product of Gaussians and δ-functions will be cumbersome in cases when stochastic variables are not independent.

In most cases, reasoning can be done using the moments (cf. Sects. 2.5); therefore, the complicated distribution is not of primary concern. The propagation of uncertainty with second moments (cf. Sect. 2.7, p. 40) only relies on the covariance matrices, not on their inverses, and can be derived usng what is called the *moment generating function* (Papoulis, 1965), which is also defined for generalized probability density functions. Thus uncertainty propagation can also be performed in mixed cases.

Similar reasoning can be used to allow random variables with zero weights $1/\sigma^2$, or infinite variance, or, more general, singular weight matrices $W = \Sigma^{-1}$ (Dempster, 1969).

2.4.5 Chi-Square Distribution

A random variable $\underline{y}$ is χ^2_n-distributed with n degrees of freedom,

$$\underline{y} \sim \chi^2_n\,, \quad \text{or} \quad \underline{y} \sim \chi^2(n), \tag{2.65}$$

if it has the density function

$$p_y(y, n) = \frac{y^{(n/2)-1} \mathrm{e}^{-y/2}}{2^{n/2}\Gamma\left(\frac{n}{2}\right)}\,, \qquad n \in \mathbb{N}\,, \qquad y > 0 \tag{2.66}$$

with the Gamma function $\Gamma(\cdot)$ (cf. Koch, 1999, Sect. 2.6.1). This distribution is used for testing quadratic forms. In particular, the sum

$$\underline{y} = \sum_{i=1}^{n} \underline{z}_i^2 \tag{2.67}$$

of n independent random variables $\underline{z}_i$, which follow a standard normal distribution ($\underline{z}_i \sim \mathcal{N}(0,1)$), is χ_n^2 distributed. For $n = 2$, we obtain the exponential distribution

$$p_y(y,2) = \frac{1}{2}\mathrm{e}^{-y/2} \qquad y \geq 0\,. \tag{2.68}$$

noncentral χ'^2 distribution

Given the n mutually independent random variables which follow noncentral normal distributions $\underline{z}_i \sim \mathcal{N}(\mu_i, 1)$, then the random variable

$$\underline{y} = \sum_{i=1}^{n} \underline{z}_i^2 \sim \chi_d'^2(\delta^2) \quad \text{with} \quad \underline{z}_i \sim \mathcal{N}(\mu_i, 1) \tag{2.69}$$

has a *noncentral chi-square distribution* $\chi_n'^2(\delta)$ with n degrees of freedom and noncentrality parameter $\delta^2 = \sum_{i=1}^{n} \mu_i^2$.

Sometimes we need the distribution of the square root $\underline{s} = \sqrt{\underline{y}}$ and thus of the length $\underline{s} = |\underline{\boldsymbol{x}}|$ of a random vector $\underline{\boldsymbol{x}} \sim \mathcal{N}(\mathbf{0}, I_n)$. The resulting distribution is the χ distribution, having density

Exercise 2.28

χ distribution

$$p_s(s,n) = \frac{2^{1-n/2} s^{n-1} \mathrm{e}^{-s^2/2}}{\Gamma(n/2)}\,. \tag{2.70}$$

2.4.6 Wishart Distribution

A symmetric positive definite $p \times p$ matrix $\underline{V}$ is Wishart distributed, $\mathcal{W}(n, \Sigma)$, with n degrees of freedom and matrix parameter Σ if its density function is (cf. Koch, 1999, Sect. 2.8.1)

$$p_W(V|n,\Sigma) = k_W \cdot |V|^{(n-p-1)/2} \mathrm{e}^{-\mathrm{tr}\left(\Sigma^{-1}V\right)/2}, n \in \mathbb{N}\,,\ |V| > 0\,,\ |\Sigma| > 0 \tag{2.71}$$

with some normalization constant k_W. This distribution is useful for evaluating empirical covariance matrices. Let N mutually independent random vectors $\underline{\boldsymbol{x}}_n$ of length p be given which follow a multivariate central normal distribution, $\underline{\boldsymbol{x}}_n \sim \mathcal{N}(\mathbf{0}, \Sigma)$. Then the matrix

$$\underline{V} = \sum_{n=1}^{N} \underline{\boldsymbol{x}}_n \underline{\boldsymbol{x}}_n^{\mathsf{T}} \sim \mathcal{W}(n, \Sigma) \tag{2.72}$$

follows a Wishart distribution. For $\Sigma = 1$ the Wishart distribution reduces to the χ^2 distribution.

Exercise 2.29

2.4.7 Fisher Distribution

A random variable $\underline{F}$ is Fisher-distributed or F-distributed,

$$\underline{F} \sim \mathcal{F}(m,n), \tag{2.73}$$

with m and n degrees of freedom if its density is

$$p_F(x|m,n) = k_F \cdot s(x) \cdot x^{\frac{m}{2}-1}(mx+n)^{-\frac{m+n}{2}} \tag{2.74}$$

with the step function $s(x)$ and a normalization constant k_F.

If two independent random variables $\underline{y}_1$ and $\underline{y}_2$ are χ^2 distributed, namely

$$\underline{y}_1 \sim \chi^2_m \qquad \underline{y}_2 \sim \chi^2_n\,, \tag{2.75}$$

then the random variable

$$\underline{F} = \frac{\underline{y}_1/m}{\underline{y}_2/n} \sim \mathcal{F}(m,n) \tag{2.76}$$

is Fisher distributed with (m,n) degrees of freedom. This distribution is used for testing results of estimation processes.

2.4.8 Student's t-Distribution

A random variable is t-distributed,

$$\underline{t} \sim t(n), \tag{2.77}$$

with n degrees of freedom, if its density is given by

$$p_t(x|n) = k_t \cdot \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}\,, \tag{2.78}$$

with some normalization constant k_t. If two independent random variables $\underline{z}$ and $\underline{y}$ are distributed according to

$$\underline{z} \sim \mathcal{N}(0,1) \qquad \underline{y} \sim \chi^2_n\,, \tag{2.79}$$

the random variable

$$\underline{t} = \frac{\underline{z}}{\sqrt{\underline{y}/n}} \sim t(n) \qquad n \in \mathbb{N} \tag{2.80}$$

follows Student's t-distribution with n degrees of freedom. This distribution may be used for testing residuals of observations after parameter estimation.

The relationships among the different distributions is given in Fig. 2.7. The normal distribution $\mathcal{N}$ is a special case of Student's t_n distribution and of the χ^2_m distribution, which themselves are special cases of the Fisher $\mathcal{F}_{m,n}$ distribution, obtained by setting one or both parameters to infinity.

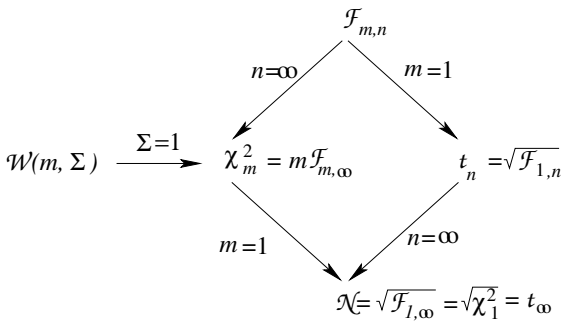

Fig. 2.7 Fisher's $\mathcal{F}_{m,n}$ and Wishart distribution $\mathcal{W}(m,\Sigma)$ and its specializations: χ^2_m, Student's t_n and normal distribution $\mathcal{N}(0,1)$. For example, taking the square root of a random variable, which is $\mathcal{F}_{1,n}$ distributed can be shown to be t_n-distributed

2.5 Moments

Moments are used to characterize probability distributions. They are mathematically equivalent to moments in physics, if the probability density function is interpreted as a mass density function.

2.5.1 General Moments

With the density functions $p_x(x)$ or $p_{xy}(x, y)$, *general moments* are defined as

$$m_r = \int_{x=-\infty}^{+\infty} x^r p_x(x)\,\mathrm{d}x \qquad r \geq 0 \tag{2.81}$$

or

$$m_{r,s} = \int_{x=-\infty}^{+\infty}\int_{y=-\infty}^{+\infty} x^r y^s p_{xy}(x, y)\,\mathrm{d}x\,\mathrm{d}y \qquad r, s \geq 0\,. \tag{2.82}$$

The values m_k and $m_{r,k-r}$, with $r \leq k$, are called *kth-order moments*. For discrete random variables with probabilities $P_x(\underline{x} = x)$ and $P_{xy}(\underline{x} = x, \underline{y} = y)$, general moments are defined as

$$m_r = \sum_{i=1}^{\infty} x_i^r P_x(\underline{x} = x_i) \qquad r \geq 0 \tag{2.83}$$

or

$$m_{r,s} = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} x_i^r y_j^s P_{xy}(\underline{x} = x_i, \underline{y} = y_j)\,\mathrm{d}x\mathrm{d}y \qquad r, s \geq 0\,. \tag{2.84}$$

We will restrict the derivations to continuous variables. The moment of the order zero is always 1. The moments m_1 or $m_{1,0}$ and $m_{0,1}$ are the *mean values* or the expected values $\mathbb{E}(\underline{x})$,

$$\mu_x \doteq m_1 = \int x p_x(x)\,\mathrm{d}x\,, \tag{2.85}$$

or

$$\mu_x \doteq m_{1,0} = \int x p_{xy}(x, y)\,\mathrm{d}x\,\mathrm{d}y\,, \tag{2.86}$$

$$\mu_y \doteq m_{0,1} = \int y p_{xy}(x, y)\,\mathrm{d}x\,\mathrm{d}y\,, \tag{2.87}$$

respectively, omitting the boundaries of the integrals.

The higher-order moments can be interpreted more easily if they refer to the mean values.

2.5.2 Central Moments

The *central moments* are defined as[2]

$$\mu_r = \int (x - \mu_x)^r p_x(x)\, \mathrm{d}x \tag{2.88}$$

and, for random d-vectors,

$$\mu_{r,s} = \int (x - \mu_x)^r (y - \mu_y)^s p_{xy}(x, y)\, \mathrm{d}x\, \mathrm{d}y\,. \tag{2.89}$$

In general, we have

$$\mu_0 = 1 \qquad \mu_1 = 0 \qquad \mu_{0,0} = 1 \qquad \mu_{1,0} = \mu_{0,1} = 0\,. \tag{2.90}$$

The central moments of a random variable yield their *variance*,

$$\sigma_x^2 \doteq \mu_2 = \int (x - \mu_x)^2 p_x(x)\, \mathrm{d}x\,, \tag{2.91}$$

$$\sigma_x^2 \doteq \mu_{2,0} = \int (x - \mu_x)^2 p_{xy}(x, y)\, \mathrm{d}x\, \mathrm{d}y\,, \tag{2.92}$$

and

$$\sigma_y^2 \doteq \mu_{0,2} = \int (y - \mu_y)^2 p_{xy}(x, y)\, \mathrm{d}x\, \mathrm{d}y\,. \tag{2.93}$$

We can easily show that the following relation holds, which in physics is called *Steiner's theorem*: *Steiner's theorem*

$$\mu_2 = m_2 - m_1^2 \qquad \text{or} \qquad \sigma_x^2 = m_2 - \mu_x^2\,. \tag{2.94}$$

Therefore, the central moments can be easily derived from the noncentral moments. The positive square root of the variance is called the *standard deviation*,

$$\sigma_x = +\sqrt{\sigma_x^2}, \tag{2.95}$$

of the random variable $\underline{x}$. The mixed second central moment of two random variables is their *covariance*

$$\sigma_{xy} \doteq \mu_{1,1} = \int (x - \mu_x)(y - \mu_y) p_{xy}(x, y)\, \mathrm{d}x\, \mathrm{d}y\,. \tag{2.96}$$

As it is difficult to interpret, it is usually related to the standard deviations σ_x and σ_y via the correlation coefficient (2.55) by

$$\sigma_{xy} = \rho_{xy} \sigma_x \sigma_y\,. \tag{2.97}$$

The second central moments of a vector $\underline{\boldsymbol{x}}$ of several random variables $\underline{\boldsymbol{x}} = [\underline{x}_i]$ usually are collected in its covariance matrix

$$\boldsymbol{\Sigma}_{xx} = [\sigma_{x_i x_j}]\,. \tag{2.98}$$

Similarly, the covariances $\sigma_{x_i y_j}$ of the random variables collected in two vectors $\underline{\boldsymbol{x}} = [\underline{x}_i]$ and $\underline{\boldsymbol{y}} = [\underline{y}_j]$ are collected in their covariance matrix

$$\boldsymbol{\Sigma}_{xy} = [\sigma_{x_i y_j}]\,. \tag{2.99}$$

Due to the symmetry of covariance matrices we have

[2] Not to be confused with the mean value μ_x.

$$\Sigma_{xy} = \Sigma_{yx}^{\mathsf{T}} \,. \tag{2.100}$$

With the diagonal matrices

$$S_x = \mathrm{Diag}([\sigma_{x_i}) \qquad S_y = \mathrm{Diag}([\sigma_{y_j}]) \tag{2.101}$$

containing the standard deviations, we can also express the covariance matrix as

$$\Sigma_{xy} = S_x R_{xy} S_y \tag{2.102}$$

using the correlation matrix

$$R_{xy} = [\rho_{x_i y_j}] = \left[\frac{\sigma_{x_i y_j}}{\sigma_{x_i} \sigma_{y_j}} \right] . \tag{2.103}$$

In the case of two random variables $\underline{x}$ and $\underline{y}$ we have their covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix} = \begin{bmatrix} \sigma_x & 0 \\ 0 & \sigma_y \end{bmatrix} \begin{bmatrix} 1 & \rho_{xy} \\ \rho_{xy} & 1 \end{bmatrix} \begin{bmatrix} \sigma_x & 0 \\ 0 & \sigma_y \end{bmatrix} . \tag{2.104}$$

We can show that covariance matrices always are positive semidefinite and the correlation coefficients ρ_{ij} always lie in $[-1, +1]$.

expectation $\mathbb{E}(\cdot)$

We use the expectation operator or mean operator $\mathbb{E}(\cdot)$ as an abbreviation. It yields the mean value of a random variable $\underline{x}$ or of a random vector $\underline{\boldsymbol{x}}$,

$$\mathbb{E}(\underline{x}) = \int_{x=-\infty}^{\infty} x p_x(x) \, \mathrm{d}x \tag{2.105}$$

and, for a d-vector $\underline{\boldsymbol{x}}$,

$$\mathbb{E}(\underline{x}) = \int_{x=-\infty}^{\infty} \boldsymbol{x} p_x(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \,. \tag{2.106}$$

The kth moments therefore are the expected or mean values of the kth power of the random variable,

$$m_k = \mathbb{E}(\underline{x}^k) \qquad m_{r,s} = \mathbb{E}(\underline{x}^r \underline{y}^s) \quad \text{with} \quad k = r + s \,. \tag{2.107}$$

The central moments thus are the expected mean values of the kth power of the difference of the random variable and its *expected* or *mean value*,

$$\mu_k = \mathbb{E}([\underline{x} - \mu_x]^k) \qquad \mu_{r,s} = \mathbb{E}([\underline{x} - \mu_x]^r [\underline{y} - \mu_y]^s) \,. \tag{2.108}$$

linearity of $\mathbb{E}(\cdot)$

The expectation operator is linear,

$$\mathbb{E}(a\underline{x} + b) = a\mathbb{E}(\underline{x}) + b \qquad \text{or} \qquad \mathbb{E}(A\underline{\boldsymbol{x}} + \boldsymbol{b}) = A\,\mathbb{E}(\underline{\boldsymbol{x}}) + \boldsymbol{b} \,, \tag{2.109}$$

which results from the linearity of the integration, a property which we often use.

Based on the expectation operator we also can define the *dispersion operator* $\mathbb{D}(\cdot)$ or $\mathbb{V}(\cdot)$ and the *covariance operator* $\mathbb{C}\mathrm{ov}(\cdot, \cdot)$, which operates on one or two vectors of random variables, respectively. The dispersion operator leads to the variance–covariance matrix of a random variable:

variance $\mathbb{V}(\cdot)$
dispersion $\mathbb{D}(\cdot)$

$$\mathbb{D}(\underline{\boldsymbol{x}}) = \mathbb{V}(\underline{\boldsymbol{x}}) = \Sigma_{xx} = \mathbb{E}[\{\underline{\boldsymbol{x}} - \mathbb{E}(\underline{\boldsymbol{x}})\}\{\underline{\boldsymbol{x}} - \mathbb{E}(\underline{\boldsymbol{x}})\}^{\mathsf{T}}] \,. \tag{2.110}$$

covariance $\mathbb{C}ov(\cdot, \cdot)$

The covariance operator leads to the covariance matrix of two random variables:

$$\mathbb{C}\mathrm{ov}(\underline{\boldsymbol{x}}, \underline{\boldsymbol{y}}) = \Sigma_{xy} = \mathbb{E}\left[\{\underline{\boldsymbol{x}} - \mathbb{E}(\underline{\boldsymbol{x}})\}\{\underline{\boldsymbol{y}} - \mathbb{E}(\underline{\boldsymbol{y}})\}^{\mathsf{T}}\right] = \Sigma_{yx}^{\mathsf{T}} = \mathbb{C}\mathrm{ov}(\underline{\boldsymbol{y}}, \underline{\boldsymbol{x}})^{\mathsf{T}} \,, \tag{2.111}$$

thus

$$\mathbb{D}(\underline{\boldsymbol{x}}) = \mathbb{V}(\underline{\boldsymbol{x}}) = \mathbb{C}\mathrm{ov}(\underline{\boldsymbol{x}}, \underline{\boldsymbol{x}})\,. \tag{2.112}$$

Observe the convention for scalar random variables x_i and y_j:

$$\Sigma_{x_i x_i} = \sigma^2_{x_i} \qquad \Sigma_{x_i y_j} = \sigma_{x_i, y_j}\,. \tag{2.113}$$

For single variables, the dispersion operator is often replaced by the variance operator, e.g., $\mathbb{V}(\underline{x}) = \sigma^2_x$.

2.5.3 Moments of Normally Distributed Variables

A variable following a one-dimensional normal distribution $\mathcal{N}(\mu, \sigma^2)$ has the first moments,

$$m_0 = 1\,, \quad m_1 = \mu\,, \quad m_2 = \mu^2 + \sigma^2\,, \quad m_3 = \mu^3 + 3\mu\sigma^2 \tag{2.114}$$

and

$$m_4 = \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4 \tag{2.115}$$

and the corresponding central moments

$$\mu_0 = 1\,, \quad \mu_1 = 0\,, \quad \mu_2 = \sigma^2\,, \quad \mu_3 = 0\,, \quad \mu_4 = 3\sigma^4\,. \tag{2.116}$$

In general, the odd central moments are zero due to the symmetry of the density function. The even central moments, $\mu_{2k}, k = 0, 1, \ldots$, of the normal distribution with density $g(x \mid \mu, \sigma^2)$ only depend on the variance

$$\mu_{2k} = \int (x-\mu)^{2k} g(x \mid \mu, \sigma^2)\, \mathrm{d}x = 1 \cdot 3 \cdot \ldots \cdot (2k-1)\sigma^{2k}\,. \tag{2.117}$$

The parameters μ and σ^2 of the one-dimensional normal distribution are the mean and the variance. The two parameters $\boldsymbol{\mu}$ and Σ of the multi-dimensional normal distribution are the mean vector and the covariance matrix.

The second (central) moment of a multi-dimensional normal distribution is the covariance matrix Σ. It exists even if the covariance matrix is singular and the density function is not a proper function.

2.5.4 Moments of the Uniform Distribution

The moments of the uniform distribution $\mathcal{U}(a, b)$ are

$$m_k = \frac{1}{k+1} \frac{b^{k+1} - a^{k+1}}{b-a}\,. \tag{2.118}$$

We obtain the even central moments $\mu_0 = 1$ and

$$\mu_2 = \sigma^2 = \frac{1}{12}(b-a)^2 \qquad \mu_4 = \frac{1}{80}(b-a)^4\,. \tag{2.119}$$

Thus, the standard deviation of the rounding error, modelled as $\underline{r} \sim \mathcal{U}\left(-\frac{1}{2}, \frac{1}{2}\right)$, is *rounding error*

$$\sigma_r = \sqrt{1/12} \approx 0.28 \tag{2.120}$$

of the last and rounded digit.

2.6 Quantiles of a Distribution

We are often interested in the value x such that the value of the cumulative distribution $P_x(x) = P(\underline{x} < x)$ is a prespecified probability α

$$P_x(x) = \int_{t=-\infty}^{x} p_x(t)\,\mathrm{d}t = \alpha\,. \tag{2.121}$$

This *α-quantile* can be determined using the *inverse cumulative distribution*

$$x = P_x^{-1}(\alpha)\,. \tag{2.122}$$

If the random variable follows a certain distribution, e.g. $\underline{x} \sim \mathcal{F}_{m,n}$, the α-quantile can be written as $x = F_{m,n;\alpha}$.

median

The *median* is the 0.5-quantile or 50th percentile

$$\mathrm{med}(x) = P_x^{-1}(0.5)\,. \tag{2.123}$$

For normally distributed random variables, it coincides with the mean, thus $N_{\mu_x,\sigma_x^2;0.5} = \mathrm{med}(x) = \mu_x$.

median absolute difference

Instead of the standard deviation, it is also possible to use the *median of the absolute differences* (MAD) from the median to characterize the spread of the random variable. It is given by

$$\mathrm{MAD}_x = \mathrm{med}(|x - \mathrm{med}(x)|)\,. \tag{2.124}$$

For normally distributed random variables, it is related to the standard deviation by

$$\mathrm{MAD}_x = \Phi^{-1}(0.75)\,\sigma_x \approx 0.6745\,\sigma_x \tag{2.125}$$

and

$$\sigma_x = \frac{1}{\Phi^{-1}(0.75)}\,\mathrm{MAD}_x \approx 1.4826\,\mathrm{MAD}_x\,, \tag{2.126}$$

(Fig. 2.5, p. 30, right).

2.7 Functions of Random Variables

Propagation of uncertainty can be formalized as follows: Given one or several random variables collected in the random vector $\underline{\boldsymbol{x}}$, together with its probability density function $p_x(\boldsymbol{x})$, and a function $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$, derive the probability density function of the random vector $\underline{\boldsymbol{y}}$.

There are several methods for solving this problem (cf. Papoulis and Pillai, 2002). We want to present two important cases with one and two random variables having arbitrary distribution and then discuss linear and nonlinear functions of Gaussian variables.

2.7.1 Transformation of a Random Variable

We first discuss the case of a monotonically increasing function $\underline{y} = f(\underline{x})$ of a single variable $\underline{x}$ with its given probability density function $p_x(x)$. The unknown probability density function of the random variable $\underline{y}$ is $p_y(y)$.

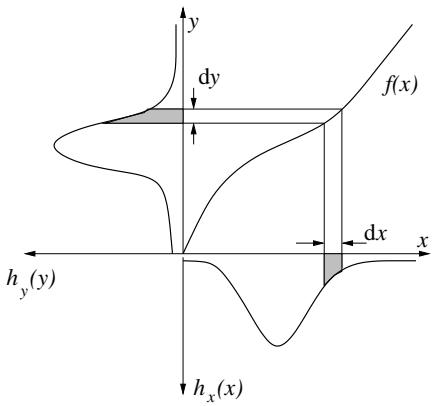

Fig. 2.8 Transformation of a random variable $\underline{x}$ with a monotonic function $y = f(x)$

With Fig. 2.8 we have $p_y(y)\,\mathrm{d}y = p_x(x)\,\mathrm{d}x$ as $P(\underline{y} \in [y, y+\mathrm{d}y]) = P(\underline{x} \in [x, x+\mathrm{d}x])$ for differential $\mathrm{d}x$ and $\mathrm{d}y$. Thus, with monotonic $f(x)$, we obtain

$$p_y(y) = \frac{p_x(x)}{\left|\frac{\mathrm{d}y}{\mathrm{d}x}\right|} = \frac{p_x(x)}{|f'(x)|}\,. \tag{2.127}$$

With the inverse function $x = f^{-1}(y)$, we finally obtain the density $p_y(y)$ of $\underline{y}$ as a function of y,

$$p_y(y) = \frac{p_x\left(f^{-1}(y)\right)}{\left|f'\left(f^{-1}(y)\right)\right|}\,. \tag{2.128}$$

This result generalizes to vector-valued variables (cf. Papoulis and Pillai, 2002, p. 142). Exercise 2.28

Example 2.7.5: ***Linear transformation of a random variable.*** For the linear transformation $y = f(x) = k + mx$, we use the first derivative $f'(x) = m$ and the inverse function

$$f^{-1}(y) = \frac{y-k}{m}$$

to obtain the density

$$p_y(y) = \frac{p_x\left(\frac{y-k}{m}\right)}{|m|}\,. \tag{2.129}$$

Obviously, the original density function $p_x(x)$ is translated by k and scaled by m in the y- and p_y-directions in order to obtain the area 1 under $p_y(y)$.

A Gaussian random variable $\underline{x} \sim \mathcal{N}(\mu, \sigma^2)$ thus can be transformed into a normalized Gaussian random variable $\underline{z} = \mathcal{N}(0,1)$ by

$$z = \frac{x-\mu}{\sigma}\,. \tag{2.130}$$

This can be generalized to a normally distributed random d-vector $\underline{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The vector *whitening*

$$\underline{\boldsymbol{z}} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d) \tag{2.131}$$

follows a normalized multivariate normal distribution. The inverse square root of the matrix $\boldsymbol{\Sigma}$ with eigenvalue decomposition $\boldsymbol{R}\boldsymbol{\Lambda}\boldsymbol{R}^\mathsf{T}$ can be determined by $\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{R}\mathrm{Diag}([1/\sqrt{\lambda_i}])\boldsymbol{R}^\mathsf{T}$. As a vector whose elements $\underline{z}_i \sim \mathcal{N}(0,1)$ are mutually independent with zero mean is called *white*, the operation (2.131) is called *whitening*. ◇

2.7.2 Distribution of the Sum of Two Random Variables

The density of the sum $\underline{z} = \underline{x} + \underline{y}$ of two independent random variables with densities $p_x(x)$ and $p_y(y)$ is

$$p_z(z) = \int p_x(z-y)p_y(y)\,\mathrm{d}y \tag{2.132}$$

$$p_z = p_x * p_y \tag{2.133}$$

and is thus identical to the *convolution* $p_x * p_y$ of the two densities p_x and p_y (Castleman, 1996).

In many cases, we have several random variables $\underline{x}_i$ which follow a joint normal distribution and which are possibly mutually correlated, $\underline{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{\mu}_x, \Sigma_{xx})$. We are interested in the distribution of new random variables $\underline{\boldsymbol{y}} = \boldsymbol{f}(\underline{\boldsymbol{x}}) = [f_i(\underline{\boldsymbol{x}})]$. Due to the nonlinearity of the functions f_i, the resulting density $p_y(\boldsymbol{y})$ is complicated.

2.7.3 Variance Propagation of Linear Functions

Probability functions often are smooth and thus may be locally approximated by a linear function. Moreover, the relative precision of the quantities involved (the random variables $\underline{\boldsymbol{x}}$) is high; thus, their standard deviations are small compared to the curvature of the functions. Under these conditions, we may approximate the resulting distribution by a normal distribution and characterize it by its first two moments, the mean and the covariance matrix.

We first give the distribution of linear functions, for which the variance propagation follows.

Given random variables $\underline{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{\mu}_x, \Sigma_{xx})$ and the linear function $\underline{\boldsymbol{y}} = A\underline{\boldsymbol{x}} + \boldsymbol{b}$, the random vector $\underline{\boldsymbol{y}}$ is normally distributed as

$$\underline{\boldsymbol{y}} \sim \mathcal{N}(A\boldsymbol{\mu}_x + \boldsymbol{b}, A\Sigma_{xx}A^\mathsf{T})\,, \tag{2.134}$$

or

$$\mathbb{E}(\underline{\boldsymbol{y}}) = A\mathbb{E}(\underline{\boldsymbol{x}}) + \boldsymbol{b}\,, \qquad \mathbb{D}(\underline{\boldsymbol{y}}) = A\mathbb{D}(\underline{\boldsymbol{x}})A^\mathsf{T}\,. \tag{2.135}$$

The proof for the preservation of the distribution uses the result of the transformation of random variables.

The proof for the first two moments uses the linearity of the expectation operator, which allows us to exchange the expectation and matrix multiplication $\mathbb{E}(\underline{\boldsymbol{y}}) = \mathbb{E}(A\underline{\boldsymbol{x}} + \boldsymbol{b}) = A\mathbb{E}(\underline{\boldsymbol{x}}) + \boldsymbol{b} = A\boldsymbol{\mu}_x + \boldsymbol{b}$ with a similar proof for the second central moments.

Comments:

- As the variance $\mathbb{V}(\underline{y}_i) = \sigma^2_{y_i}$ of an arbitrary element $\underline{y}_i$ for arbitrary matrices A needs to be nonnegative, the covariance matrix Σ_{xx} needs to be positive semi-definite.
- Though the density function of the normal distribution is not defined for singular covariance matrices, the probability function exists. Variance propagation uses only the moments, so it is allowed for singular covariance matrices as well. If A does not have full rank, then Σ_{yy} is singular.
- The proof only uses the moments. It is thus valid for arbitrary distributions $\mathcal{M}_x(\boldsymbol{\mu}_x, \Sigma_{xx})$ for which we only use the first two moments, $\boldsymbol{\mu}_x$ and Σ_{xx}. Therefore, we have the following law of *variance propagation*:

variance propagation

$$\underline{\boldsymbol{x}} \sim \mathcal{M}_x(\boldsymbol{\mu}_x, \Sigma_{xx}) \quad \text{and} \quad \underline{\boldsymbol{y}} = A\underline{\boldsymbol{x}} + \boldsymbol{b} \quad \rightarrow \quad \underline{\boldsymbol{y}} \sim \mathcal{M}_y(A\boldsymbol{\mu}_x + \boldsymbol{b}, A\Sigma_{xx}A^\mathsf{T})\,. \tag{2.136}$$

- The inverse W_{xx} of a regular covariance matrix Σ_{xx} is sometimes called a *weight matrix* or the *precision matrix* (cf. Bishop, 2006),

  weight matrix, precision matrix

  $$W_{xx} = \Sigma_{xx}^{-1}\,, \tag{2.137}$$

  as random variables with smaller variances have higher weights and higher precision when performing an estimation (Sect. 4.1.4, p. 79).
  If A is invertible, we also have a propagation law for weight matrices,

  $$W_{yy} = A^{-1} W_{xx} A^{-\mathsf{T}}\,. \tag{2.138}$$

- We can transfer the result to linear functions of random matrices. Given the random matrix $\underline{X} \sim \mathcal{M}(\mathbb{E}(\underline{X}), \mathbb{D}(\mathrm{vec}\underline{X}))$ and the linear function $\underline{Y} = A\underline{X}B + C$, the random matrix $\underline{Y}$ is normally distributed since

  $$\underline{Y} \sim \mathcal{M}(A\mathbb{E}(\underline{X})B + C, (B^{\mathsf{T}} \otimes A)\Sigma_{xx}(B^{\mathsf{T}} \otimes A)^{\mathsf{T}})\,. \tag{2.139}$$

  Using the vectors $\underline{\boldsymbol{x}} = \mathrm{vec}\underline{X}$ and $\underline{\boldsymbol{y}} = \mathrm{vec}\underline{Y}$ this result immediately follows from the vectorized function $\underline{\boldsymbol{y}} = (B^{\mathsf{T}} \otimes A)\underline{\boldsymbol{x}} + \mathrm{vec}C$ (cf. (A.95), p. 775).

2.7.4 Variance Propagation of Nonlinear Functions

In the case of nonlinear functions $\underline{\boldsymbol{y}} = \boldsymbol{f}(\underline{\boldsymbol{x}})$, we first perform a Taylor series expansion,

$$\underline{\boldsymbol{y}} = \boldsymbol{y}^{(0)} + \mathrm{d}\underline{\boldsymbol{y}} = \boldsymbol{f}(\boldsymbol{x}^{(0)}) + J\mathrm{d}\underline{\boldsymbol{x}} + O(|\mathrm{d}\underline{\boldsymbol{x}}|^2), \tag{2.140}$$

with the Jacobian

$$J = [J_{ij}] = \left[\frac{\partial f_i(\boldsymbol{x})}{\partial x_j}\right]\Bigg|_{x=x^{(0)}}\,, \tag{2.141}$$

where – to simplify notation – the subscript $x = x^{(0)}$ refers to the vector $\boldsymbol{x}$. If we use $\boldsymbol{x}^{(0)} = \boldsymbol{\mu}_x$ with $\boldsymbol{y}^{(0)} = \boldsymbol{f}(\boldsymbol{x}^{(0)})$, we obtain

$$\mathrm{d}\underline{\boldsymbol{y}} = J\,\mathrm{d}\underline{\boldsymbol{x}}, \tag{2.142}$$

and therefore in a first-order approximation

$$\mathbb{E}(\underline{\boldsymbol{y}}) \approx \boldsymbol{\mu}_y^{(1)} = \boldsymbol{f}(\boldsymbol{\mu}_x)\,, \qquad \mathbb{D}(\underline{\boldsymbol{y}}) \approx \Sigma_{yy}^{(1)} = J\Sigma_{xx}J^{\mathsf{T}} \tag{2.143}$$

since, up to a first-order approximation,

$$\Sigma_{yy} = \Sigma_{\mathrm{d}y\,\mathrm{d}y} \tag{2.144}$$

due to $\underline{\boldsymbol{y}} \approx \boldsymbol{y}^{(0)} + \mathrm{d}\underline{\boldsymbol{y}}$.

It can be shown that with relative errors $r_{x_j} = \sigma_{x_j}/\mu_{x_j}$ of the variables $\underline{x}_i$, the error in the standard deviations σ_{y_j} due to linearization is less than $r_{x_j}\sigma_{y_i}$, and is thus negligible in most practical applications; cf. Sect. 2.7.6, p. 44.

2.7.5 Implicit Variance Propagation

If we have an implicit relation

$$\boldsymbol{f}(\underline{\boldsymbol{x}}, \underline{\boldsymbol{y}}) = \boldsymbol{0} \tag{2.145}$$

between two stochastic variables $\underline{\boldsymbol{x}}$ and $\underline{\boldsymbol{y}}$, the variance propagation can be performed with the Jacobians

$$A = \left.\frac{\partial \boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})}{\partial \boldsymbol{x}}\right|_{x=\mu_x, y=\mu_y} \qquad B = \left.\frac{\partial \boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})}{\partial \boldsymbol{y}}\right|_{x=\mu_x, y=\mu_y} \tag{2.146}$$

if B is invertible. From $\mathrm{d}\boldsymbol{f} = A\,\mathrm{d}\boldsymbol{x} + B\,\mathrm{d}\boldsymbol{y} = \boldsymbol{0}$ we obtain $\mathrm{d}\underline{\boldsymbol{y}} = -B^{-1}A\,\mathrm{d}\underline{\boldsymbol{x}}$ with given Σ_{xx}, again, in a first-order approximation,

$$\Sigma_{yy} = B^{-1}A\Sigma_{xx}A^{\mathsf{T}}B^{-\mathsf{T}}\,. \tag{2.147}$$

This allows the derivation of the covariance matrix of $\underline{\boldsymbol{y}}$ even if the procedure for deriving $\boldsymbol{y}$ from $\boldsymbol{x}$ is very complicated.

2.7.6 Bias Induced by Linearization

bias: deviation from the true value

Moment propagation (2.143) of nonlinear functions using only the first-order Taylor series of the nonlinear function leads to a systematic deviation from the *true value*, also called *bias*. Analysing higher-order terms yields expressions for the bias due to linearization.

For a scalar function $y = f(x)$ of a scalar x, it is based on the Taylor expansion of the stochastic variable at $f(\mu_x)$,

$$\begin{aligned}\underline{y} = f(\underline{x}) = f(\mu_x) + f'(\mu_x)(\underline{x}-\mu_x) + \frac{1}{2}f''(\mu_x)(\underline{x}-\mu_x)^2 \\ + \frac{1}{6}f'''(\mu_x)(\underline{x}-\mu_x)^3 + \frac{1}{24}f^{(4)}(\mu_x)(\underline{x}-\mu_x)^4 + O((\underline{x}-\mu_x)^n)\,.\end{aligned} \tag{2.148}$$

We therefore obtain the following result: if the density function of a stochastic variable $\underline{x}$ is symmetrical, the mean for $\underline{y} = f(\underline{x})$ can be shown to be

$$\mathbb{E}(\underline{y}) = \mu_y = f(\mu_x) + \frac{1}{2}f''(\mu_x)\sigma_x^2 + \frac{1}{24}f^{(4)}(\mu_x)\mu_{4_x} + O(f^{(n)}, m_n) \qquad n > 4. \tag{2.149}$$

For normally distributed variables we take its central fourth moment $\mu_{4_x} = 3\sigma_x^4$. Using the expression $\mathbb{V}(\underline{y}) = \mathbb{E}(\underline{y}^2) - [\mathbb{E}(\underline{y})]^2$ from (2.94), p. 37 we can derive a similar expression for the variance. Restricting to even moments up to the fourth-order for Gaussian variables, we have

Exercise 2.30

$$\mathbb{V}(\underline{y}) = \left[\sigma_y^{(2)}\right]^2 = f'^2(\mu_x)\,\sigma_x^2 + \left(f'(\mu_x)f'''(\mu_x) + \frac{1}{2}f''^2(\mu_x)\right)\sigma_x^4 + O(f^{(n)}, m_n)\,. \tag{2.150}$$

Obviously the bias, i.e., the second term, depends on the variance and the higher-order derivatives: the larger the variance and the higher the curvature or the third derivative, the higher the bias. Higher-order terms again depend on derivatives and moments of order higher than 4.

expectation of function of stochastic vector

For a stochastic vector $\underline{\boldsymbol{x}}$ with symmetrical density function, the mean of the scalar function $y = f(\boldsymbol{x})$ can be shown to be

Exercise 2.31

$$\mathbb{E}(\underline{y}) = \mu_y^{(2)} = f(\boldsymbol{\mu}_x) + \frac{1}{2}\mathrm{trace}(H|_{x=\mu_x}\cdot\Sigma_{xx}) + O(f^{(n)}, m_n), \quad n \geq 3\,, \tag{2.151}$$

with the Hessian matrix $H = (\partial f^2/\partial x_i \partial x_j)$ of the function $f(\boldsymbol{x})$. This is a generalization of (2.149).

We now discuss two cases in more detail which regularly occur in geometric reasoning, the bias of a product and the bias of normalizing a vector to length 1.

Bias of a Product. The product $z = xy$ of two random variables is part of all geometric constructions when using *homogeneous coordinates* for representing geometric entities. For the product

$$\underline{z} = \underline{x}\,\underline{y} \tag{2.152}$$

of two possibly correlated normal random variables

$$\begin{bmatrix} \underline{x} \\ \underline{y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \rho_{xy}\sigma_x\sigma_y \\ \rho_{xy}\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix} \right), \tag{2.153}$$

we obtain the first and second approximation for the mean value Exercise 2.32

$$\mu_z^{[1]} = \mu_x\mu_y \qquad \mu_z^{[2]} = \mu_z^{[1]} + \rho_{xy}\sigma_x\sigma_y\,. \tag{2.154}$$

Thus we obtain the *bias of the mean*,

$$b_{\mu_z} \doteq \mu_z^{[2]} - \mu_z^{[1]} = \sigma_{xy} = \rho_{xy}\sigma_x\sigma_y, \tag{2.155}$$

and the *relative bias of the mean* of the product,

$$r_{\mu_z} \doteq \frac{b_{\mu_z}}{\mu_z} = \rho_{xy}\frac{\sigma_x}{\mu_x}\frac{\sigma_y}{\mu_y}\,. \tag{2.156}$$

The relative bias of the mean is the product of the relative accuracies σ_x/μ_x and σ_y/μ_y multiplied with the correlation coefficient. The bias is zero if the random variables are uncorrelated, which is often the case when constructing a geometric entity from two others. The proof of (2.154), p. 45 uses Exercise 2.33

$$\mathbb{E}((\underline{x} - \mu_x)^2(\underline{y} - \mu_y)^2) = (1 + 2\rho_{xy})\sigma_x^2\sigma_y^2\,. \tag{2.157}$$

Similarly, we have the first- and second-order approximation for the standard deviation, Exercise 2.34

$$\sigma_z^{[1]} = \mu_y^2\sigma_x^2 + \mu_x^2\sigma_y^2 + 2\mu_x\mu_y\sigma_{xy} \qquad \sigma_z^{[2]} = \sigma_z^{[1]} + (1 + \rho_{xy}^2)\sigma_x^2\sigma_y^2\,. \tag{2.158}$$

The *bias of the variance* is

$$b_{\sigma_z^2} = \sigma_z^{2[2]} - \sigma_z^{2[1]} = \sigma_x^2\sigma_y^2 + \sigma_{xy}^2 = (1 + \rho_{xy}^2)\sigma_x^2\sigma_y^2\,, \tag{2.159}$$

and therefore the *relative bias of the variance*,

$$r_{\sigma_z^2} = \frac{b_{\sigma_z^2}}{\sigma_z^2} = \frac{(1 + \rho_{xy}^2)\sigma_x^2\sigma_y^2}{\mu_y^2\sigma_x^2 + \mu_x^2\sigma_y^2 + 2\mu_x\mu_y\sigma_{xy}}, \tag{2.160}$$

does not vanish for uncorrelated random variables.

If the variables are uncorrelated and have the same relative precision, i.e., $\sigma_x/\mu_x \approx \sigma_y/\mu_y \approx \sigma/\mu$, we obtain the relative bias

$$r_{\sigma_z^2} = \frac{b_{\sigma_z^2}}{\sigma_z^2} \approx \frac{1}{2}\left(\frac{\sigma}{\mu}\right)^2. \tag{2.161}$$

Thus, the relative bias $r_{\sigma_z^2}$ of the variance is approximately half of the square of the relative precision σ/μ.

Bias of Normalization. The normalization of an n-vector $\mathbf{x}$ to unit length, which we will apply to homogeneous coordinates regularly (Sect. 5.1, p. 195), is given by

$$\mathbf{x}^s = \frac{\mathbf{x}}{|\mathbf{x}|} \qquad \text{or} \qquad \mathbf{x}_i^s = \frac{x_i}{|\mathbf{x}|}\,. \tag{2.162}$$

We assume $\underline{\boldsymbol{x}}$ has covariance matrix $\Sigma_{xx} = \sigma_x^2 I_n$. This leads to the following expression for the mean when taking terms up to the fourth-order into account:

Exercise 2.35

$$\mathbb{E}(\underline{\mathbf{x}}^s) = \frac{\boldsymbol{\mu}_x}{|\boldsymbol{\mu}_x|} \left(1 - \frac{1}{2} \frac{\sigma_x^2}{|\boldsymbol{\mu}_x|^2}\right) . \tag{2.163}$$

Here too, the relative bias, since it is identical to the bias, is approximately half of the square of the relative accuracy.

The bias of the variance behaves in a similar manner as for the product of two entities: the relative bias of the variance follows quadratically with the relative precision of the given entities; cf. (2.161).

In nearly all cases which are practically relevant when geometrically analysing images, the relative precision results from the observation process in images, which is below one pixel (see the following example). Even for wide-angle cameras, the focal length is far beyond 100 pixels. The directional uncertainty is therefore much better than one percent. As a consequence, the relative bias when determining the mean value or the variance using only the first-order approximation is significantly smaller than 0.01%.

2.7.7 On the Mean and the Variance of Ratios

Care has to be taken when deriving Euclidean coordinates, $\boldsymbol{x}$, from homogeneous ones, $\mathbf{x}$, e.g., using the ratios

$$x = \frac{u}{w} \qquad y = \frac{v}{w} \tag{2.164}$$

if the denominator w is uncertain. If $\underline{w} \sim \mathcal{N}(\mu_w, \sigma_w^2)$, the mean and the variance of $\underline{x}$ and $\underline{y}$ are not defined (cf. Hartley and Zisserman, 2000, App. 3). The reason is that with a possibly very small probability the denominator w will be zero; thus, the variable x will be infinite, making the integral $\mu_x = \int_{-\infty}^{\infty} x p(x) \mathrm{d}x$ vanish.

Exercise 2.36

However, the first-order approximation for deriving the mean $\mu_x = \mu_u/\mu_w$ and the variance is still useful due to the practical procedure of preprocessing the observed data $\mathbf{x}$: they are usually checked for outliers, and only the inliers are used in further processing. This preprocessing limits the range of possible random perturbations for the inlying observations, and would make it necessary to work with a distribution with limited support, say $\pm 4\sigma_w$:

$$\underline{w} \mid \text{inlier} \sim p_{w|inlier}(w|\text{inlier}) = \begin{cases} k \cdot g(w \mid \mu_w, \sigma_w^2), & \text{if } w \in [\mu_w - 4\sigma_w, \mu_w + 4\sigma_w] \\ 0, & \text{else} \end{cases} \tag{2.165}$$

with an adequate normalization constant k for the truncated Gaussian density g. This distribution has approximately the same first and second moments as the corresponding Gaussian but does not cause infinite mean or variance if $|\mu_w|$ is far enough from zero, i.e., $|\mu_w| > 4\sigma_w$. Therefore, the classical determination of the mean and the variance by using variance propagation is sufficiently accurate.

In order to be able to handle outliers as well, we model the causing gross error as a shift b_w of the mean,

$$\underline{w} \mid \text{outlier} \sim p_{w|inlier}(w - b_w) , \tag{2.166}$$

which also allows variance propagation and is consistent with the model of classical hypothesis testing (Sect. 3.1.1, p. 62), which is the basis for outlier detection, e.g., in a RANSAC procedure (Sect. 4.7.7, p. 153).

We therefore recommend using variance propagation based only on the linearized relations. The example on p. 48 supports the recommendation.

2.7.8 Unscented Transformation

Classical variance propagation of nonlinear functions only uses the first-order terms of the Taylor series. The bias induced by omitting higher-order terms in many practical cases is irrelevant.

We now discuss a method which uses terms up to the fourth-order and in many cases yields results which are accurate up to the second-order. It is called *unscented transformation* (cf. Julier and Uhlmann, 1997).

It is based on the idea of representing the distribution of the given random N-vector $\underline{\boldsymbol{x}}$ by $2N+1$ well-selected points $\boldsymbol{x}_i$ and of deriving the weighted mean vector and the covariance matrix from the nonlinearly transformed points $\boldsymbol{y}_n = \boldsymbol{f}(\boldsymbol{x}_n)$.

The selected points depend on the square root

$$\mathsf{S}_{xx} = \sqrt{\Sigma_{xx}} = [\boldsymbol{s}_n]\,, \qquad \Sigma_{xx} = \mathsf{S}_{xx}\mathsf{S}_{xx}^{\mathsf{T}} \tag{2.167}$$

of the covariance matrix of the given random variable. Its columns are $\boldsymbol{s}_n$. For numerical reasons, S_{xx} is best determined by Cholesky decomposition (Rhudy et al., 2011). Now we have

$$\Sigma_{xx} = [\boldsymbol{s}_1, ..., \boldsymbol{s}_n, ...\boldsymbol{s}_N] \begin{bmatrix} \boldsymbol{s}_1^{\mathsf{T}} \\ ... \\ \boldsymbol{s}_n^{\mathsf{T}} \\ ... \\ \boldsymbol{s}_N^{\mathsf{T}} \end{bmatrix} = \sum_{n=1}^{N} \boldsymbol{s}_n \boldsymbol{s}_n^{\mathsf{T}}\,. \tag{2.168}$$

The $2N+1$ points $\boldsymbol{x}_n$ and their weights w_n then are:

$$\begin{aligned} \boldsymbol{x}_1 &= \boldsymbol{\mu}_x\,, & w_1 &= \frac{\kappa}{N+\kappa} \\ \boldsymbol{x}_n &= \boldsymbol{\mu}_x + \sqrt{N+\kappa}\,\boldsymbol{s}_n\,, & w_n &= \frac{1}{2(N+\kappa)} \qquad n = 2, ..., N+1 \\ \boldsymbol{x}_{n+N} &= \boldsymbol{\mu}_x - \sqrt{N+\kappa}\,\boldsymbol{s}_n\,, & w_n &= \frac{1}{2(N+\kappa)} \qquad n = N+2, ..., 2N+1\,. \end{aligned} \tag{2.169}$$

They depend on a free parameter κ. The weights add to 1. For Gaussian random variables, we best use

$$\kappa = 3 - N \tag{2.170}$$

in order to obtain minimum bias. As a result, some of the weights may be negative.

Determining the mean and covariance matrix of $\underline{\boldsymbol{y}}$ is performed in three steps:

1. transforming the points

$$\boldsymbol{y}_n = \boldsymbol{f}(\boldsymbol{x}_n) \qquad n = 1, ..., 2N+1\,, \tag{2.171}$$

2. determining the mean vector

$$\boldsymbol{\mu}_y = \sum_{n=1}^{2N+1} w_n \boldsymbol{y}_n\,, \tag{2.172}$$

and

3. determining the covariance matrix

$$\Sigma_{yy} = \sum_{n=1}^{2N+1} w_n (\boldsymbol{y}_n - \boldsymbol{\mu}_y)(\boldsymbol{y}_n - \boldsymbol{\mu}_y)^{\mathsf{T}} = \left(\sum_{n=1}^{2N+1} w_n \boldsymbol{y}_n \boldsymbol{y}_n^{\mathsf{T}} \right) - \boldsymbol{\mu}_y \boldsymbol{\mu}_y^{\mathsf{T}}\,. \tag{2.173}$$

Example 2.7.6: ***Unscented transformation of a linear function.*** In the case of a linear function $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{a}$, we obtain the same mean and covariance matrix as with the classical variance propagation.

Proof: The mean value $\boldsymbol{\mu}_y$ is obviously identical to $\boldsymbol{f}(\boldsymbol{\mu}_x)$. For the covariance matrix, we use the transformed points $\boldsymbol{y}_1 - \boldsymbol{\mu}_y = \boldsymbol{0}$ and $\boldsymbol{y}_n - \boldsymbol{\mu}_y = \pm\sqrt{N+\kappa}\,\boldsymbol{A}\boldsymbol{s}_n$. Then (2.173) yields

$$\boldsymbol{\Sigma}_{yy} = \sum_{n=1}^{N} \frac{1}{2(N+\kappa)} \left((\sqrt{N+\kappa})^2 \boldsymbol{A}\boldsymbol{s}_n \boldsymbol{s}_n^{\mathsf{T}} \boldsymbol{A}^{\mathsf{T}}) + (\sqrt{N+\kappa})^2 (-\boldsymbol{A}\boldsymbol{s}_n)(-\boldsymbol{s}_n^{\mathsf{T}} \boldsymbol{A}^{\mathsf{T}}) \right) = \boldsymbol{A}\boldsymbol{\Sigma}_{xx}\boldsymbol{A}^{\mathsf{T}} .$$

◇

Example 2.7.7: ***Square of a standard Gaussian random variable.*** Here we have $\underline{x} \sim \mathcal{N}(0,1)$ and the function $y = f(x) = x^2$. The mean and the variance can be derived from the general properties of the χ^2 distribution. For the sum $\underline{z} \sim \chi^2_N$ of N squared independent random variables $\underline{u}_n \sim \mathcal{N}(0,1)$, the mean and variance are

$$\mathbb{E}(\underline{z}^2) = N \qquad \mathbb{D}(\underline{z}^2) = 2N . \tag{2.174}$$

In our special case, $n = 1$, the mean is

$$\mathbb{E}(\underline{x}^2) = 1 , \qquad \mathbb{D}(\underline{x}^2) = 2 . \tag{2.175}$$

The classical variance propagation leads to completely wrong results $\mu_y^{(1)} = 0$ and $\sigma_y^{(1)} = 0$, as $y(0) = y'(0) = 0$.

With the unscented transformation, with $N = 1$ we use the $2N + 1 = 3$ points and weights:

$$x_1 = 0 , \quad w_1 = \frac{2}{3} , \qquad x_2 = \sqrt{3} , \quad w_2 = \frac{1}{6} , \qquad x_3 = -\sqrt{3} , \quad w_3 = \frac{1}{6} . \tag{2.176}$$

Therefore we obtain

1. the transformed points $y_1 = 0$, $y_2 = y_3 = 3$,
2. the weighted mean
$$\mu_y = \frac{2}{3} \cdot 0 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 3 = 1 , \tag{2.177}$$
3. the weighted sum of the squares $\sum_{n=1}^{3} w_n y_n^2 = 3$ and therefore the variance
$$\sigma_y^2 = \sum_{n=1}^{3} w_n y_n^2 - \mu_y^2 = 2 . \tag{2.178}$$

Comparison with (2.175) shows that the unscented transformation in this highly nonlinear case yields the correct result. ◇

2.8 Stochastic Processes

In this section we discuss sequences of random variables and their statistical properties. We will use such processes for modelling surface profiles in Chap. 16, p. 727. We address two types of models: (1) using (auto-) covariance functions,[3] which specify the process by its second-order statistics, and (2) using autoregressive processes, which refer to the first-order statistics. Both models allow the generation of sample processes and the estimation of the underlying parameters. They differ in the efficiency for interpolation and the ease of generalizing the concept from one to two dimensions.

2.8.1 Notion of a Stochastic Process

stochastic process

Following the introduction of random variables in Sect. 2.3, p. 24, a stochastic process associates to a certain outcome $s \in S$ of an experiment a function $\underline{x}(t, s)$ depending on the independent variable t (Papoulis and Pillai, 2002): The function

[3] This is in contrast to crosscovariance functions between two different processes.

$$\underline{x}(t) : S \to \mathbb{F} \qquad \underline{x}(t) = \underline{x}(t, s) \tag{2.179}$$

is called a stochastic process. The range $\mathbb{F}$ of functions is to be specified. This notion naturally can be generalized to more functions of more than one variable if the scalar t is replaced by a d-dimensional vector. We start with functions of one variable t as they naturally occur as time series.

Depending on whether we fix t or s, we can interpret $\underline{x}(t, s)$ as

1. a stochastic process $\underline{x}(t, s)$, if t and s are variables ,
2. a sampled function $x(t)$, if s is fixed,
3. a random variable $\underline{x}(s)$, if t is fixed and s is variable, and
4. a sampled value x at time t, if s and t are fixed.

A stochastic process is completely specified by the distribution function

$$P(x_1, ..., x_n; t_1, ..., t_n) = P(\underline{x}(t_1) \leq x_1, ..., \underline{x}(t_n) \leq x_n) \tag{2.180}$$

for arbitrary n and $t_1, ..., t_n$. A stochastic process is called *stationary in the strict sense* if the distribution function is invariant to a shift of the parameters t_n by a common delay. *strict stationarity*

We distinguish between continuous and discrete processes, depending on whether t is taken from a continuous domain $\mathcal{D} \subseteq \mathbb{R}$ or whether t is taken from a discrete domain, e.g., $\mathcal{D} \subseteq \mathbb{Z}$. If a process is discrete, we use n as an independent variable and write

$$\underline{x}(n) = \underline{x}(n, s) , \qquad n \in \mathbb{Z} \tag{2.181}$$

where x depends on a discrete parameter n. Such processes can be interpreted as sequences of random variables, e.g., $\underline{x}(n), n = 1, ..., N$.

Furthermore, we only address Gaussian processes. They are fully characterized by their first and second moments

$$\mu_x(t) = \mathbb{E}(\underline{x}(t)) \quad \text{and} \quad \sigma_{xx'}(t, t') = \mathbb{C}\text{ov}(\underline{x}(t), \underline{x}(t')) \tag{2.182}$$

$$\mu_x(n) = \mathbb{E}(\underline{x}(n)) \quad \text{and} \quad \sigma_{xx'}(n, n') = \mathbb{C}\text{ov}(\underline{x}(n), \underline{x}(n')) . \tag{2.183}$$

In the following paragraphs we refer to continuous and discrete processes using t as an argument.

A stochastic process is called *weakly stationary* if the first and second moments do not depend on time. Then we have $\mu_x(t) = \mu_x(t')$ or *weak stationarity*

$$\mu_x = \mathbb{E}_x(\underline{x}(t)) = \int x\, p(x, t)\, \mathrm{d}x \qquad \text{for all } t \tag{2.184}$$

and $\sigma(t + u, t' + u) = \sigma(t, t')$. With the difference between two variables, which is called the *lag*, *lag*

$$d = t' - t, \tag{2.185}$$

we obtain

$$\sigma_{xx'}(d) = \sigma_{xx'}(t, t + d) = \sigma_{xx'}(-d) , \tag{2.186}$$

the last relation resulting from the symmetry of the covariance of two random variables. The function $\sigma_{xx'}(d)$ is the *covariance function* of the stationary process and often written as *covariance function*

$$C_{xx}(d) = \mathbb{C}\text{ov}(\underline{x}(t), \underline{x}(t + d)) . \tag{2.187}$$

A stationary stochastic process is therefore characterized by its mean μ_x and its covariance function $C_{xx}(d)$.

We first discuss continuous processes specified by their covariance function, and then a special class of models which define the sequence of the random variables recursively.

2.8.2 Continuous Gaussian Processes

A stationary continuous Gaussian process is characterized by the mean value μ_x and the covariance function $C_{xx}(d)$. We discuss the main properties of covariance functions.

Stationary One-Dimensional Gaussian Processes. The covariance function C_{xx} needs to guarantee that, for any i, the vector $\boldsymbol{x} = [x(t_i)], i = 1, ..., I$, the covariance matrix

$$\begin{aligned}\Sigma_{xx} = \mathbb{D}(\underline{\boldsymbol{x}}) &= \begin{bmatrix} \mathbb{C}\mathrm{ov}\,(x(t_1), x(t_1)) & \dots & \mathbb{C}\mathrm{ov}\,(x(t_1), x(t_i)) & \dots & \mathbb{C}\mathrm{ov}\,(x(t_1), x(t_I)) \\ \dots & \dots & \dots & \dots & \dots \\ \mathbb{C}\mathrm{ov}\,(x(t_i), x(t_1)) & \dots & \mathbb{C}\mathrm{ov}\,(x(t_i), x(t_i)) & \dots & \mathbb{C}\mathrm{ov}\,(x(t_i), x(t_I)) \\ \dots & \dots & \dots & \dots & \dots \\ \mathbb{C}\mathrm{ov}\,(x(t_I), x(t_1)) & \dots & \mathbb{C}\mathrm{ov}\,(x(t_I), x(t_i)) & \dots & \mathbb{C}\mathrm{ov}\,(x(t_I), x(t_I)) \end{bmatrix} \\ &= \begin{bmatrix} C_{xx}\,(0)) & \dots & C_{xx}\,(t_1 - t_i) & \dots & C_{xx}\,(t_1 - t_I) \\ \dots & \dots & \dots & \dots & \dots \\ C_{xx}\,(t_i - t_1) & \dots & C_{xx}\,(0) & \dots & C_{xx}\,(t_i - t_I) \\ \dots & \dots & \dots & \dots & \dots \\ C_{xx}\,(t_I - t_1) & \dots & C_{xx}\,(t_I - t_i) & \dots & C_{xx}\,(0) \end{bmatrix} \end{aligned} \tag{2.188}$$

positive semi-definite and positive definite functions

is positive semi-definite. This can be achieved if we choose a *positive semi-definite function*. Following Bochner's theorem (cf. Rasmussen and Williams, 2005, Sect. 4.2.1), a positive definite function is a function whose Fourier transform is positive, or which can be written as

$$C_{xx}(d) = \sum_{k=0}^{\infty} c_k \cos(2\pi k d) \tag{2.189}$$

with

$$\sigma_x^2 = \sum_{k=0}^{\infty} c_k < \infty \quad \text{and} \quad c_k > 0, \text{ for all } k\,. \tag{2.190}$$

If the coefficients fulfil $c_k \geq 0$, the function is called positive semi-definite. Observe that the diagonal elements of the covariance matrix are identical to the variance of the process: $C_{xx}(0) = \sigma_x^2$. Similarly we have positive semi-definite correlation functions using (2.103), p. 38,

$$R_{xx}(d) = \frac{C_{xx}(d)}{C_{xx}(0)} = \frac{C_{xx}(d)}{\sigma_x^2}\,. \tag{2.191}$$

Examples of correlation functions are

$$R_1(d) = \begin{cases} 1, & \text{if } d = 0 \\ 0, & \text{else} \end{cases} \tag{2.192}$$

$$R_2(d) = \exp\left(-\frac{|d|}{|d_0|}\right) \tag{2.193}$$

$$R_3(d) = \exp\left(-\frac{1}{2}\left(\frac{d}{d_0}\right)^2\right) \tag{2.194}$$

$$R_4(d) = \frac{1}{1 + \left(\frac{d}{d_0}\right)^2} \tag{2.195}$$

with some reference distance d_0.

Linear combinations $h(d) = af(d) + bg(d)$ with positive coefficients a and b and products $h(d) = f(d)g(d)$ of two positive functions $f(d)$ and $g(d)$ again are positive definite functions.

Figure 2.9 shows three samples of a Gaussian process $x(t_k), k = 1, 2, 3$. The standard deviation of the processes is $\sigma_x = 1$. The covariance function is $C_{xx}(d) = \exp\left(-\frac{1}{2}(d/20)^2\right)$, cf. R_3 in (2.194). The method for generating such sequences is given in Sect. 2.9, p. 55.

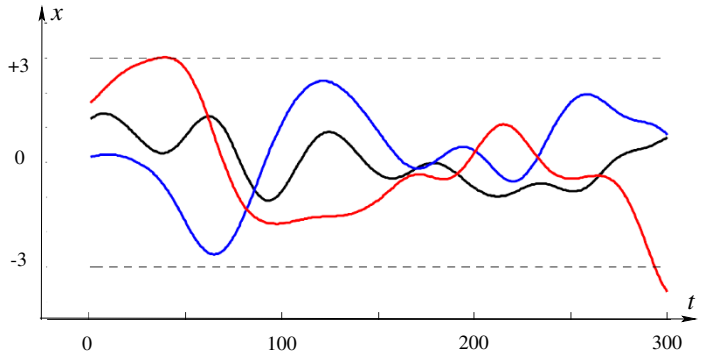

Fig. 2.9 Three samples of size 300 of a Gaussian process with mean 0, standard deviation $\sigma_x = 1$, and correlation function $R_3(d)$ with $d_0 = 20$

Homogeneous and Isotropic Higher Dimensional Gaussian Processes. The concept of stationary Gaussian processes can be generalized to functions depending on two or more variables, collected in a vector, say $\boldsymbol{u}$. They usually are applied to represent spatial stochastic processes. We refer to a two-dimensional stochastic process $\underline{x}(\boldsymbol{u}, s)$ in the following. It will be used to describe the random nature of surfaces, where x represents the height and $\boldsymbol{u} = [u, v]$ the position.

For spatial processes the concept of invariance to translation is called *homogeneity*, which is equivalent to the notion of *stationarity* for time processes. Moreover, the characteristics of spatial processes may be also invariant to rotation. A higher dimensional stochastic process is called *isotropic* if the covariance between two values $\underline{x}(\boldsymbol{u}_1)$ and $\underline{x}(\boldsymbol{u}_2)$ does not depend on a rotation of the coordinate system: $\mathbb{C}\text{ov}(\underline{x}(\boldsymbol{u}_1), \underline{x}(\boldsymbol{u}_2)) = \mathbb{C}\text{ov}(R\underline{x}(\boldsymbol{u}_1), R\underline{x}(\boldsymbol{u}_2))$ for an arbitrary rotation matrix R.

homogeneous stochastic process

isotropic stochastic process

Now, homogeneous and isotropic Gaussian processes can again be characterized by their mean μ_x and their covariance function

$$C_{xx}(d(\boldsymbol{u}, \boldsymbol{u}')) = \mathbb{C}\text{ov}(x(\boldsymbol{u}), x(\boldsymbol{u}')) \tag{2.196}$$

where the distance $d = d(\boldsymbol{u}, \boldsymbol{u}') = |\boldsymbol{u}' - \boldsymbol{u}|$ is the Euclidean distance between the positions $\boldsymbol{u}$ and $\boldsymbol{u}'$. Again, an arbitrary covariance matrix Σ_{xx} must be positive semi-definite.

Remark: If the distance $d = |\boldsymbol{u}' - \boldsymbol{u}|$ is replaced by a weighted distance, say $d = \sqrt{(\boldsymbol{u}' - \boldsymbol{u})^\mathsf{T} W (\boldsymbol{u}' - \boldsymbol{u})}$, with a constant positive definite matrix W, the stochastic process still is homogeneous, but anisotropic. Generalizing the concept to nonhomogeneous anisotropic processes is out of the scope of this book. ◇

Representing stochastic processes using covariance functions can be seen as characterizing the second moments of vectors of random variables, where the index refers to a parameter, say t, of a continuous or discrete domain. This has the advantage of generalizing the concept to more dimensions. Next we discuss a class of models for stochastic processes which are based on a generative model for the process itself, which has the advantage of leading to more efficient computational schemes.

2.8.3 Autoregressive Processes

An autoregressive model AR(P) of order P is characterized by P parameters $a_p, p = 1, ..., P$, and a variance σ_e^2. It uses a sequence $\underline{e}_n \sim \mathcal{M}(0, \sigma_e)$ of identically and independently distributed (iid) random variables. This sequence controls the stochastic development of the stochastic process $\underline{x}_n$; therefore, it is often called the driving process. Starting from a set of P random variables $\underline{x}_n$, with $\mathbb{E}(\underline{x}_n) = 0$, the elements $\underline{x}_n, n > P$, of the random sequence linearly and deterministically depend on the previous P values, $\underline{x}_{n-p}$ of the sequence and the nth element, $\underline{e}_n$, of the driving process, in the following manner:

$$\underline{x}_n = \sum_{p=1}^{P} a_p \, \underline{x}_{n-p} + \underline{e}_n \,, \quad \underline{e}_n \sim \mathcal{M}\left(0, \sigma_e^2\right) \,, \quad n > P \,. \tag{2.197}$$

Since $\mathbb{E}(\underline{e}_n) = 0$, we have

$$\mathbb{E}(\underline{x}_n) = 0 \,. \tag{2.198}$$

If this condition is not fulfilled, the process model may be modified by adding the mean value c:

$$\underline{x}_n = c + \sum_{p=1}^{P} a_p (\underline{x}_{n-p} - c) + \underline{e}_n \,, \quad \underline{e}_n \sim \mathcal{M}\left(0, \sigma_e^2\right) \tag{2.199}$$

The stochastic process is stationary if the generally complex zeros of the polynomial $1 - \sum_{p=1}^{P} a_p z^p$ are outside the unit circle (cf. Box and Jenkins, 1976). We illustrate the situation for the autoregressive model AR(1).

AR(1) Processes. An AR(1) model, using $a := a_1$ for simplicity, is given by:

$$\underline{x}_n = a\underline{x}_{n-1} + \underline{e}_n \,, \quad \underline{e}_n \sim \mathcal{M}\left(0, \sigma_e^2\right) \quad \text{and} \quad |a| < 1 \,. \tag{2.200}$$

We choose the initial value $\underline{x}_0 \sim \mathcal{M}(0, 0)$ and

$$\underline{e}_1 \sim \mathcal{M}\left(0, \frac{1}{1-a^2}\sigma_e^2\right) \tag{2.201}$$

intentionally in order to obtain a stationary process, as can be seen immediately. We recursively obtain

$$\underline{x}_1 = \underline{e}_1 \qquad \sigma_{x_1}^2 = \frac{1}{1-a^2}\sigma_e^2 \tag{2.202}$$

$$\underline{x}_2 = a\underline{e}_1 + \underline{e}_2 \qquad \sigma_{x_2}^2 = \left(\frac{a^2}{1-a^2} + 1\right)\sigma_e^2 \tag{2.203}$$

$$\underline{x}_3 = a^2\underline{e}_1 + a\underline{e}_2 + \underline{e}_3 \qquad \sigma_{x_3}^2 = \left(\frac{a^4}{1-a^2} + a^2 + 1\right)\sigma_e^2 \tag{2.204}$$

$$\ldots \qquad \ldots \tag{2.205}$$

$$\underline{x}_n = a^{n-1}\underline{e}_1 + a^{n-2}\underline{e}_2 + \ldots + \underline{e}_n \qquad \sigma_{x_n}^2 = \left(\frac{a^{2(n-1)}}{1-a^2} + a^{2(n-2)} + \ldots + 1\right)\sigma_e^2 \,. \tag{2.206}$$

As can be checked easily, we therefore have

$$\sigma_x^2 = \frac{\sigma_e^2}{1-a^2} \tag{2.207}$$

independent of n. Obviously, only values $|a| < 1$ lead to stationary sequences with limited variance:

1. For $a = 0$ we have a white noise process.
2. For $a \in (0, 1)$ the process is randomly deviating from zero while keeping close to 0.
3. For $a \in (-1, 0)$ the process is oscillating while staying close to 0.
4. For $|a| > 1$ and first increment $\underline{e}_1 \sim \mathcal{M}(0, \sigma_e^2)$ the process x_n is quickly diverging with $\sigma_{x_n}^2 = (a^{2n} - 1)/(a^2 - 1)\,\sigma_e^2$.

Furthermore, from (2.202)ff. we obtain the covariance function, i.e., the covariance between neighbouring random variables $\underline{x}_n$ and $\underline{x}_{n+d}$,

$$C_{xx}(d) = \mathbb{C}\mathrm{ov}(\underline{x}_n, \underline{x}_{n+d}) = a^d\, \sigma_{x_n}^2\,, \tag{2.208}$$

which is an exponential function of the lag d. Thus the correlation (2.55), p. 31 between neighbouring variables

$$\rho_d = \rho_{x_n, x_{n+d}} = a^d \tag{2.209}$$

decays exponentially with the distance d for $|a| < 1$. The covariance matrix of a sequence $\{x_n\}$ with N values, collected in the N-vector $\underline{\boldsymbol{x}}$, therefore is

$$\mathbb{D}(\underline{\boldsymbol{x}}) = \frac{\sigma_e^2}{1 - a^2} \begin{bmatrix} 1 & a & a^2 & \dots & a^{N-2} & a^{N-1} \\ a & 1 & a & \dots & a^{N-3} & a^{N-2} \\ a^2 & a & 1 & \dots & a^{N-4} & a^{N-3} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a^{N-2} & a^{N-3} & a^{N-4} & \dots & 1 & a \\ a^{N-1} & a^{N-2} & a^{N-3} & \dots & a & 1 \end{bmatrix} = \frac{\sigma_e^2}{1 - a^2} \left[a^{|i-j|} \right] . \tag{2.210}$$

This matrix has a special structure. Its off-diagonal elements only depend on the distance $|i - j|$ from the main diagonal. Such matrices are called *Toeplitz* matrices. *Toeplitz matrix*

Integrated White Noise Processes. For $a = 1$ we obtain a special process: It is a summed white noise process, often called an *integrated white noise process*,

$$\underline{x}_n = \underline{x}_{n-1} + \underline{e}_n\,, \quad \mathbb{D}(\underline{e}_n) = \sigma_e^2 \tag{2.211}$$

with starting value $\underline{x}_0 = 0$. The name of this process results from the sequence

$$\underline{x}_1 = \underline{e}_1 \tag{2.212}$$
$$\underline{x}_2 = \underline{e}_1 + \underline{e}_2 \tag{2.213}$$
$$\underline{x}_3 = \underline{e}_1 + \underline{e}_2 + \underline{e}_3 \tag{2.214}$$
$$\dots = \dots \tag{2.215}$$
$$\underline{x}_n = \sum_{k=1}^{n} \underline{e}_k\,. \tag{2.216}$$

Two samples for such a process with different standard deviations of the driving noise process are given in Fig. 2.10, upper row. They are generated using a random number generator for the sequence e_k (cf. .Sect. 2.9). Rewriting the generating equation in the form

$$\underline{e}_n = \underline{x}_n - \underline{x}_{n-1} \tag{2.217}$$

reveals the driving white noise sequence $\{\underline{e}_n\}$ to represent the discrete approximation of the first derivative of the discrete function x_n. The process is slowly diverging with $\sigma_n = \sqrt{n}\,\sigma_e$. It is not a stationary process.

If we apply a second summation we arrive at the second-order autoregressive process AR(2) with coefficients $a_1 = 2$ and $a_2 = -1$, a *doubly integrated white noise process*,

$$\underline{x}_n = 2\underline{x}_{n-1} - \underline{x}_{n-2} + \underline{e}_n\,, \quad \mathbb{D}(\underline{e}_n) = \sigma_e^2 \tag{2.218}$$

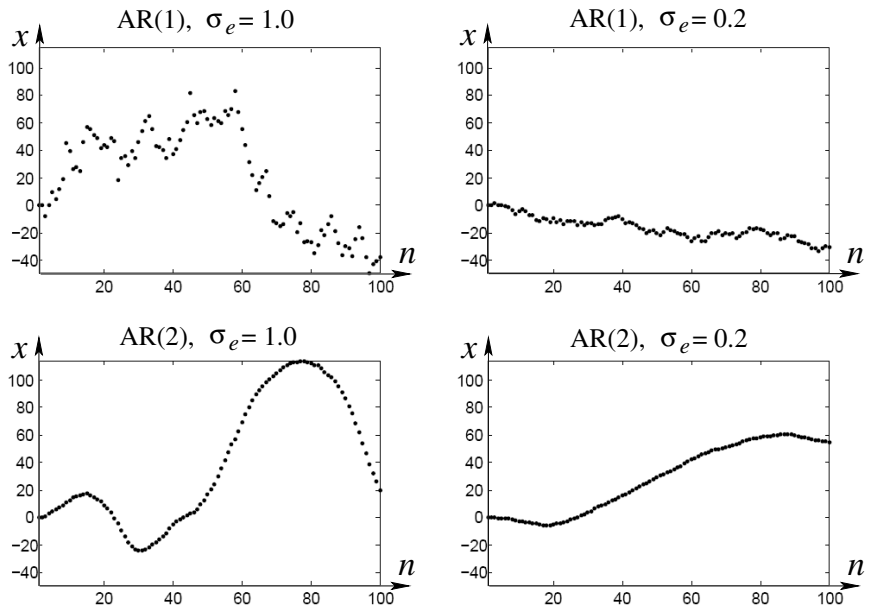

Fig. 2.10 Examples for autoregressive processes. Sequences of 100 points. Integrated and doubly integrated white noise processes (upper and lower row) with standard deviation of driving noise process $\sigma_e = 1.0$ and $\sigma_e = 0.2$ (left and right column)

with starting values values $\underline{x}_0 = \underline{x}_{-1} = 0$. Two examples for such a process are given in Fig. 2.10, lower row. Again solving for $\underline{e}_n$ yields

$$\underline{e}_n = \underline{x}_n - 2\underline{x}_{n-1} + \underline{x}_{n-2}\,. \tag{2.219}$$

Thus $\underline{e}_n$ measures the second derivative of the sequence $\underline{x}_n$ at position $n-1$. Again, as the mean value of the driving noise process $\underline{e}_n$ is zero, the variance σ_e^2 of the AR(2) process measures the smoothness of the sequence.

2.8.4 Integrated AR Processes

We have discussed two versions of an integrating process, where a white noise process drives it. This idea can be generalized to situations where the white noise process drives the first- or higher-order derivatives of the process. When the Dth derivatives of a process follow an AR(P) model, the process is called an integrated autoregressive process, and denoted by ARI(P, D).

As an example, we have an autoregressive model ARI(P,2) for the sequence of second derivatives,

$$\underline{x}_{n-1} - 2\underline{x}_n + \underline{x}_{n+1} = \sum_{p=1}^{P} a_p \underline{x}_{n-p} + \underline{e}_n\,, \tag{2.220}$$

which will turn out to be a good model for terrain profiles. Obviously, this model can be written as

$$\underline{x}_{n+1} = -(\underline{x}_{n-1} - 2\underline{x}_n) + \sum_{p=1}^{P} a_p \underline{x}_{n-p} + \underline{e}_n \tag{2.221}$$

or as an AR($P+1$)-process. It can be written as

$$\underline{x}_n = -(\underline{x}_{n-2} - 2\underline{x}_{n-1}) + \sum_{p=1}^{P} a_p \underline{x}_{n-p-1} + \bar{\underline{e}}_n \tag{2.222}$$

$$= 2\underline{x}_{n-1} + a_1\underline{x}_{n-2} + a_2\underline{x}_{n-3} + ... + a_P\underline{x}_{n-(P+1)} + \bar{\underline{e}}_n \tag{2.223}$$

$$= \sum_{q=1}^{P+1} b_q \underline{x}_{n-q} + \bar{\underline{e}}_n \tag{2.224}$$

with coefficients

$$b_1 = 2\,, \quad b_2 = a_1 - 1\,, \qquad b_q = a_{q-1} \text{ for } q = 3, ..., P+1\,, \quad \bar{\underline{e}}_n = \underline{e}_{n-1}\,. \tag{2.225}$$

2.9 Generating Random Numbers

Testing algorithms involving random variables can be based on simulated data. Here we address the generation of random variables following a certain distribution, which then can be used as input for an algorithm. Software systems provide functions to generate samples of most of the distributions given in this chapter. Visualization of the distributions can be based on scatterplots or histograms.

Take as an example a random variable $\underline{y} \sim \mathcal{N}(\mu_y, \sigma_y^2)$. We want to visualize its distribution for given μ_y and variance σ_y^2. Provided we have a routine for generating a random variable $\underline{x} \sim \mathcal{N}(0, 1)$, we can derive a sample y of a random variable $\underline{y}$ using (2.134), p. 42. We choose the linear function

$$y = \mu_y + \sigma_y\, x \tag{2.226}$$

to derive a sample y from a sample x. Repeating the generation process usually provides statistically independent samples, a property which has to be guaranteed by the random number generator. Alternatively the provided routine allows us to generate vectors or matrices of random numbers. As an example, the package MATLAB provides the function `x=randn(N,M)` to generate an $N \times M$ matrix of random variables $\underline{x}_{nm}$ which follow a standard normal distribution $\underline{x} \sim \mathcal{N}(0, 1)$.

The samples for the autoregressive processes in Fig. 2.10, p. 54 have been generated using a vector $\underline{\boldsymbol{e}}$ of normally distributed random variables $\underline{e}_n$.

A large sample of N values $\underline{x}_n$ can be taken to visualize the distribution via the histogram. The histogram takes a set of K bins $[x_k, x_{k+1})$, which are half open intervals, and counts the number N_k of samples in the bins. The bins usually are equally spaced. A useful number K for the bins is $K = \lfloor\sqrt{N}\rfloor$, as this is a balance between too narrow and too few bins. As the probability P_k that a sample value lies in a bin is $P_k = \int_{x=x_k}^{x_{k+1}} p_x(x)\mathrm{d}x$, and N_k/N is an estimate for this probability, the form of the histogram can be visually compared to the theoretical density $p_x(x)$ by overlaying the histogram by the function $N\,P_k$ using the approximation of $P(\underline{x} \in [x, x + \mathrm{d}x]) = p_x(x)\mathrm{d}x$ (cf. (2.16), p. 26, and Fig. 2.11, top right), namely

$$P_k \approx \frac{1}{2}\left(p_x(x_k) + p_x(x_{k+1})\right)\ \ (x_{k+1} - x_k)\,. \tag{2.227}$$

If we want to generate a sample of a vector of normally distributed values $\underline{\boldsymbol{y}} \sim \mathcal{N}(\boldsymbol{\mu}_y, \Sigma_{yy})$, we can proceed similarly. We start from a vector $\boldsymbol{x} = [x_n], n = 1, ..., N$, where the independent samples $\underline{x}_n \sim \mathcal{N}(0, 1)$ follow a standard normal distribution, thus $\underline{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{0}, I_N)$. We need the square root S_{yy} of the covariance matrix Σ_{yy} (cf. (2.167), p. 47). Then the linear function

$$\boldsymbol{y} = \boldsymbol{\mu}_y + S_{yy}\,\boldsymbol{x} \tag{2.228}$$

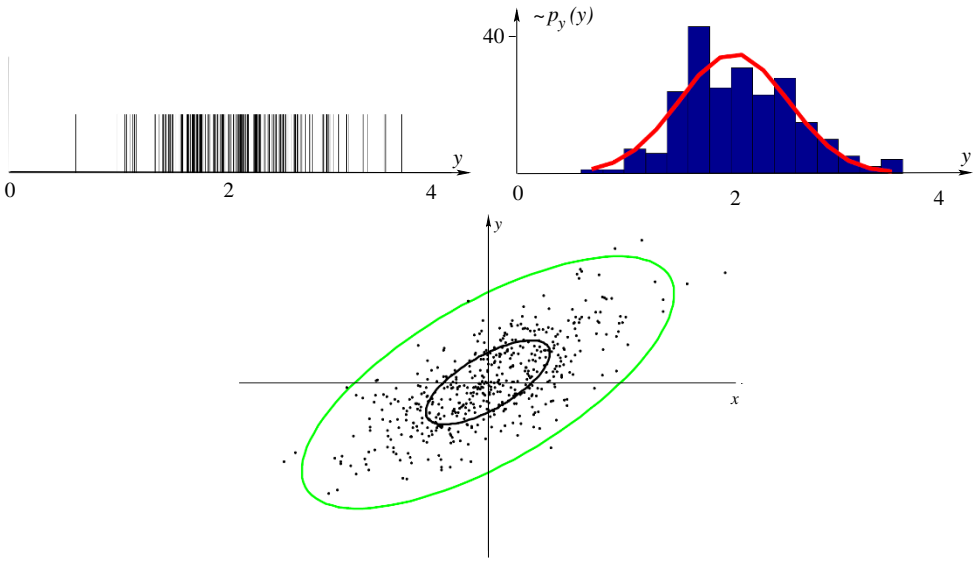

Fig. 2.11 Top row left: One-dimensional scatter plot of a sample of $N = 225$ normally distributed random variables $\underline{y} \sim \mathcal{N}(2, 0.25)$. **Top row right:** Histogram of the same sample with 15 bins, overlayed with its probability density. **Bottom:** 2D scatter plot of $N = 500$ samples of normally distributed random vectors overlayed with the standard ellipse (black) and threefold standard ellipse (green) (Fig. (2.6), p. 32). Approximately 99% of the samples lie in the threefold standard ellipse (Table 2.2, p. 32): $d = 2$, $S = 0.99$

of the sample $\boldsymbol{x}$ of the random vector $\underline{\boldsymbol{x}}$ leads to a sample vector $\boldsymbol{y}$ with distribution $\underline{\boldsymbol{y}} \sim \mathcal{N}(\boldsymbol{\mu}_y, \Sigma_{yy})$.

The Gaussian processes in Fig. 2.9, p. 51 have been realized by (1) specifying a regular sequence of $N = 300$ arguments $t = 1, ..., N$, (2) generating the $N \times N$ covariance matrix Σ_{xx} using the standard deviation $\sigma_x = 1$ and the correlation function $R_3(d)$, and (3) taking samples from a normally distributed vector $\underline{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{0}, \Sigma_{xx})$.

Exercise 2.37 Samples of other distributions can be generated using similar routines.

2.10 Exercises

The number in brackets at the beginning of each exercise indicate its difficulty, cf. Sect. 1.3.2.4, p. 16

Basics

1. (1) How could you randomly choose a month when throwing a die twice? Is the expected probability of all months the same?
2. (1) Give a probability the sun will shine tomorrow? What are the problems when giving such a number?
3. (2) Take a die and throw it repeatedly. Determine the probability of the event 1 after every sixth throw following von Mises' definition of probability. Describe how the determined probability evolves over time. When do you expect to be able to prove that the determined probability converges towards 1/6?
4. (2) You throw a die four times. What is the probability of throwing the sequence $(1, 2, 3, 4)$? What is the probability of throwing three even numbers? What is the probability of throwing 6 at least twice, if the first two throws are $(3, 6)$. What is the probability of throwing the sum 10?

5. (1) Plot the probability and the density function for throwing the numbers 1 to 6 with a die. What would change if the die did not show numbers but six different colours?
6. (2) Plot the density function of n times throwing a 6 when throwing a die $N = 3$ times. Give the density function $p(n)$ explicitly. What is the probability in this experiment of throwing a 6 at least once? Show this probability in a plot of the cumulative probability function.
7. (2) Assume the display of a range sensor can show numbers between 0.000 and 999.999. The sensor may fail, yielding an outlier. Assume the sensor shows an arbitrary number $\underline{s}$ if it fails. Describe the random variable $\underline{s}$ for the outlier. Is it a discrete or continuous random variable? How large is the difference between a discrete and a continuous model for the outlier? What is the probability that $\underline{s} \in [100, 110]$ in the discrete and the continuous model? What changes if the display shows numbers only up to one digit after the dot, i.e., in the range 0.0 to 999.0?
8. (2) Plot the density function of random variables $\underline{x}$ and $\underline{y}$ following the exponential and the Laplace distribution, respectively. Give names to the axes. Give the probability that $\underline{x} \in [-1, 2]$ and $\underline{y} \in [-1, 2]$.

Computer Experiments

9. (3) Use a program for generating M samples of a normal distribution $\mathcal{N}(0, 1)$. Determine the histogram

$$h(x_i|b) = \#(\underline{x} \in [x_i - b/2, x_i + b/2]\,, \quad x_i = ib\,, \quad b \in \mathbb{R}\,,\, i \in \mathbb{Z} \tag{2.229}$$

   from M samples. Prespecify the bin size b. Determine the probability $p(x_i|b) = h(x_i|b)/M$ that a sample falls in a certain bin centred at x_i. Overlay the plot with the density function of the normalized normal distribution $\phi(x)$. How do you need to scale the axes such that the two functions $\phi(x)$ and $p(x_i|b)$ are comparable. Vary the bin size b and the number of samples M. What would be a good bin size if M is given?
10. (2) Repeat the previous exercise for M samples y_m of a χ-square distribution with n degrees of freedom. For this generate $\underline{y}_m$ as the sum of the squares of n samples from a standard normal distribution. Also vary the degrees of freedom n. Describe the distribution for $n = 1, 2, 3$ and for large n.
11. (2) Prove that the bounding box for the standard ellipse has size $2\sigma_x \times 2\sigma_y$. *Hint:* Show the y-coordinate of the highest and lowest point of the ellipse is $\pm\sigma_y$ based on the partial derivative of $(\boldsymbol{x} - \boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) = 1$ w.r.t. x, see (2.56), p. 31.
12. (3) Generate a covariance matrix V following a Wishart distribution $\underline{\mathsf{V}} \sim \mathcal{W}(n, \mathsf{I}_2)$. Plot the standard ellipse of V. Repeat the experiment and observe the variation of V. Vary $n = 5, 10, 50$ and discuss the result.
13. (2) This and the following exercise show that it is sufficient to determine the noncentral moments of basic variables, since the central moments and moments of transformed variables linearly depending on the original variables can be expressed as functions of the noncentral moments. As an example we have the relation between the second central moment μ_2 and the moments m_1 and m_2, given by $\mu_2 = m_2 - m_1^2$. This can be generalized to higher-order moments.

    Express the third central moments of a distribution $\mu_{ij}, i + k = 3$ as a function of the third moments $m_{ij}, i + j = 3$.
14. (3) Let the moments of two variables $\underline{x}$ and $\underline{y}$ be denoted by $m_x := m_{10}$, $m_y := m_{01}$, $m_{xx} := m_{20}$, etc. Derive the central second moments m_{uu}, m_{uv}, m_{vv} of the rotated variables $\underline{u}$ and $\underline{v}$,

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}, \tag{2.230}$$

    as a function of ϕ and the noncentral moments of $\underline{x}$ and $\underline{y}$.

15. (1) Given are two correlated random variables $\underline{x}$ and $\underline{y}$ with the same standard deviation σ. Give the standard deviations and the correlation of their sum and their difference. How does the result specialize, if (a) the two random variables are uncorrelated, (b) are correlated with 100%, and (c) are correlated with minus 100%?
16. (1) Show that the correlation coefficient ρ_{xy} between two stochastic variables $\underline{x}$ and $\underline{y}$ lies in the interval $[-1, +1]$, as the covariance matrix needs to be positive semi-definite. Show that the covariance matrix is singular if and only if $\rho = \pm 1$.
17. (1) Prove $\mathbb{E}(a\underline{x} + b) = a\mathbb{E}(\underline{x}) + b$, see (2.109), p. 38.
18. (2) Given are three stochastically independent random variables, $\underline{x} \sim \mathcal{M}(3, 4)$, $\underline{y} \sim \mathcal{M}(-2, 1)$, and $\underline{z} \sim \mathcal{M}(1, 9)$.
    a. (1) Derive the mean and the standard deviation of the two functions
    $$u = 1 + 2x - y\,, \qquad v = -3 + 2y + 3z\,. \tag{2.231}$$
    b. (1) What is the correlation coefficient ρ_{uv}?
    c. (1) Let a further random variable be $\underline{w} = \underline{u} + \underline{z}$. What is the variance of $\underline{w}$ and its correlation ρ_{xw} with $\underline{x}$?
    d. (1) What is the covariance matrix $\mathbb{C}\mathrm{ov}(\underline{u}, [\underline{v}, \underline{w}]^\mathsf{T})$?
19. (2) We want to approximate the normal distribution $\mathcal{N}(\mu, \sigma^2)$ by a uniform distribution such that the mean and the variance is identical to the normal distribution. Give the parameters a and b. Especially relate the range $r = b - a$ of the uniform distribution to the standard deviation σ of the normal distribution. Compare the result to $\sigma_r = \sqrt{1/12}$, see (2.120), p. 39.
20. (1) Given a sequence $\underline{g}(i) \sim \mathcal{M}(\mu(i), \sigma^2), i = 1, 2, 3, ...$ of random variables representing a noisy sampled signal $g(t)$, its discrete derivative can be determined from $g_t(i) = (g(i + 1) - g(i - 1))/2$. Determine the standard deviation of $\underline{g}_t(i)$.
21. (3) We say a random variable $\underline{z} \sim k\chi^2_n$ follows a $k\chi^2_n$ distribution if $\underline{z}/k \sim \chi^2_n$. Given an array $\underline{g}_{ij} \sim \mathcal{M}(\mu_{ij}, \sigma^2)$ of random variables, representing a noisy sampled function $g(x, y)$, the partial derivatives can be derived from
    $$g_x(i, j) = (g(i + 1, j) - g(i - 1, j))/2\,, \quad g_y(i, j) = (g(i, j + 1) - g(i, j - 1))/2\,. \tag{2.232}$$
    Give the standard deviations of the two partial derivatives and their covariance. What is the distribution of the squared magnitude $\underline{m}^2(i, j) := |\nabla \underline{g}(i, j)|^2 = \underline{g}^2_x(i, j) + \underline{g}^2_y(i, j)$ of the gradient $\nabla g = [g_x, g_y]^\mathsf{T}$? *Hint:* Which distribution would $\underline{m}^2$ follow if the two random variables $\underline{g}_x$ and $\underline{g}_y$ were standard normally distributed?
22. (1) Let $\underline{y} \sim \chi^2_2$ be χ-square distributed with two degrees of freedom. Determine the mean μ_y. Relate the α-percentile $\chi_{2,\alpha}$ to the mean.
23. (2) Given a random variable $\underline{x} \sim \mathcal{N}(0, 1)$, show that $\underline{x}^2 \sim \chi^2_1$.
24. (2) Given the basis b of two cameras with principal distance c and the x-coordinates x' and x'' of the two image points of a scene point, its distance Z from the camera is given by
    $$Z = \frac{bc}{x'' - x'}\,. \tag{2.233}$$
    Assume the variables, namely b, c, x', and x'', are uncertain, with individual standard deviations σ_b, σ_c, $\sigma_{x'}$, and $\sigma_{x''}$, respectively, and mutually independent. Derive the standard deviation σ_Z of $\underline{Z}$. Derive the relative precision σ_Z/μ_Z of $\underline{Z}$ as a function of the relative precision of the three variables $\underline{b}$, $\underline{c}$, and $\underline{p} = \underline{x}'' - \underline{x}'$.
25. (2) Given are two points $\boldsymbol{p} = [2, 1]^\mathsf{T}$ m and $\boldsymbol{q} = [10, 9]^\mathsf{T}$ m. Their distances to an unknown point $\boldsymbol{x} = [x, y]$ are $s = 5$ m and $t = 13$ m and have standard deviation $\sigma_s = \sigma_r = 0.1$ m.

a. (1) Prove that the two intersection points of the circles around $\boldsymbol{p}$ and $\boldsymbol{q}$ are $\boldsymbol{x}_1 = [14, 6]^\mathsf{T}$ m and $\boldsymbol{x}_2 = [7, 13]^\mathsf{T}$ m.
b. (2) Derive the covariance matrix of the intersection point $\boldsymbol{x}_1$.

26. (3) Given is the function $y = f(x) = x^4 - x^3$ and the random variable $\underline{x} \sim \mathcal{N}(0, 1)$. Derive the mean and the variance of $\underline{y} = f(\underline{x})$

    a. using variance propagation,
    b. using the unscented transformation,
    c. using 10,000 samples of $\underline{x}$ as reference,

    and compare.

Proofs

27. (1) Steiner's theorem ((2.94), p. 37) relates the noncentral second and the central second moments of a variable via the mean. Generalize the theorem to multivariate variables.
28. (1) Prove the expression (2.70), p. 34 for the χ distribution. *Hint:* Apply (2.128), p. 41 to (2.66), p. 33.
29. (1) Refer to the Wishart distribution ((2.71), p. 34) and prove that for $\Sigma = 1$ and $V = y$ we obtain the χ^2 distribution ((2.66), p. 33).
30. (1) Prove the expression (2.150), p. 44 for the second-order approximation for the variance.
31. (1) Prove the expression (2.151), p. 44 for the second-order approximation of the mean of a function depending on a vector.
32. (1) Prove the first- and second-order approximation (2.154), p. 45 for the mean of a product.
33. (2) Prove the expression (2.157), p. 45 for the expectation of $(\underline{x} - \mu_x)^2(\underline{y} - \mu_y)^2$ of two correlated Gaussian variables. *Hint:* Assume $\mu_x = \mu_y = 0$.
34. (1) Prove the expression (2.158), p. 45 for the second-order approximation of the expectation of a random vector, which is normalized to length 1.
35. (1) Prove (2.163), p. 46. *Hint:* use (2.151), p. 44 for each component x_i of $\mathbf{x}$.
36. (1) Let the random variable $\underline{x} \sim \mathcal{N}(m, \sigma_x^2)$ with $m > 0$ be given. Let the derived random variable be $\underline{y} = 1/\underline{x}$.
    Using (2.149), p. 44 and (2.117), p. 39, derive a general expression for the odd moments of $\mathbb{E}(\underline{y})$. Show that the series for odd n begins with

$$\mathbb{E}\left(\frac{1}{\underline{x}}\right) = \frac{1}{\mu_x}\left(1 + \frac{\sigma_x^2}{\mu_x^2} + \frac{3\sigma_x^4}{\mu_x^4} + \frac{15\sigma_x^6}{\mu_x^6} + \ldots\right) \tag{2.234}$$

    Show that the series diverges.
37. (1) Given the cumulative distribution $P_x(x)$ of a random variable $\underline{x}$, show that the random variable $P_x^{-1}(\underline{y})$ has density $p_x(x)$ if $\underline{y}$ is uniformly distributed in the interval $[0, 1]$.

http://www.springer.com/978-3-319-11549-8

Photogrammetric Computer Vision
Statistics, Geometry, Orientation and Reconstruction
Förstner, W.; Wrobel, B.P.
2016, XVII, 816 p. 281 illus., 59 illus. in color.,
Hardcover
ISBN: 978-3-319-11549-8