Probabilistic Modeling with the Wolfram Language

Gerhard Hejc

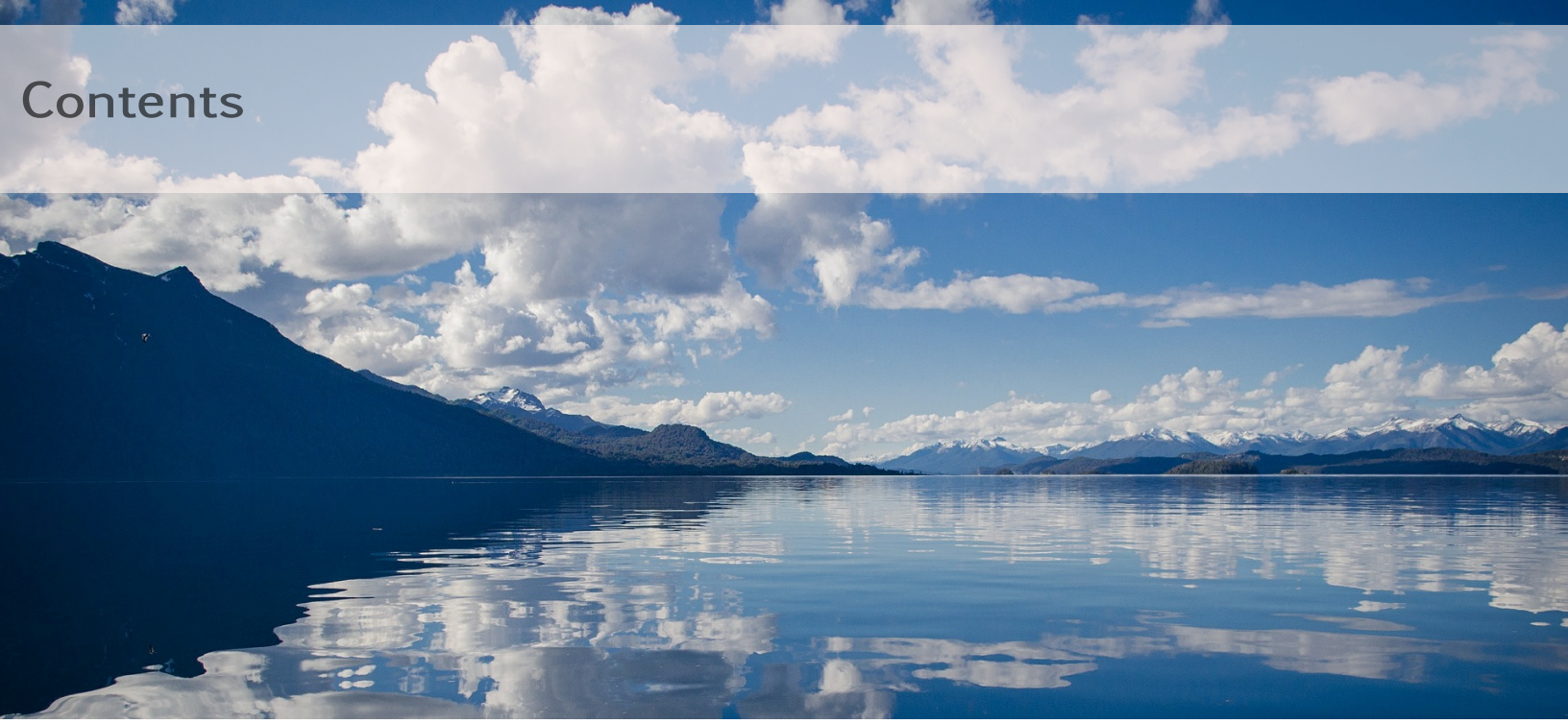
Publisher

Art. No xxxxx
ISBN xxx–xx–xxxx–xx–x
Edition 0.0

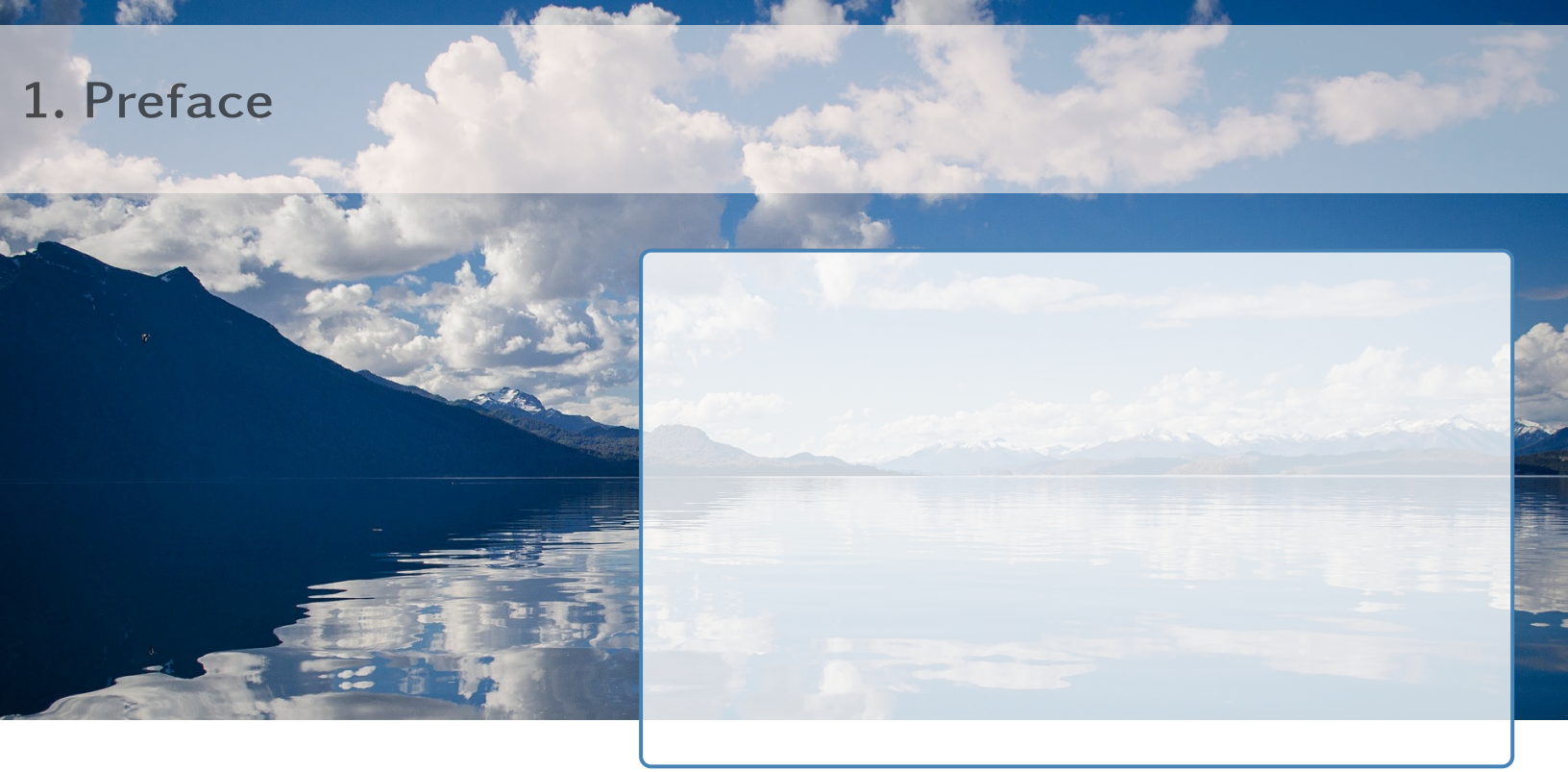
Cover design by Cover Designer

Published by Publisher
Printed in City

Contents



1. Preface



The book originally developed from talks and lecture notes given at the Fraunhofer institute in Nuremberg, Germany as an introduction to probabilistic modeling. The goal was primarily to provide the students with a minimal set of mathematical tools and concepts to do probabilistic computations by themselves and to apply it to real-world problems. This sounds easier than it actually is, because even conceptually easy problems can lead to results, which contradicts intuition like the famous Monty Hall problem.

Since then, a lot of developments in this field took place. Most of the non-trivial calculations are almost impossible to be done without a computer, so the question quickly arises what is the appropriate software tool for doing this kind of calculations. Because of the availability of a large number of good software, it is merely an author's preference towards one or the other, but in this book the Wolfram language is chosen for several reasons.

The Wolfram language has a clear and consistent design with a syntax close to the mathematical notation and has all the built-in knowledge to perform symbolic calculations where results in a closed-form are available and numerical evaluations in the case where this is not possible. Especially the Probability and Statistics part, which is used heavily in this book, has well-designed functions, which makes it easy to apply them to all kind of statistical problems. The visualization capabilities of the Wolfram language are outstanding and every figure in this book was generated with it.

This is not the first book about this topic, but builds on the work of others. There are many great books available, and these recommendations are definitely not exhaustive, but only a subjective subset:

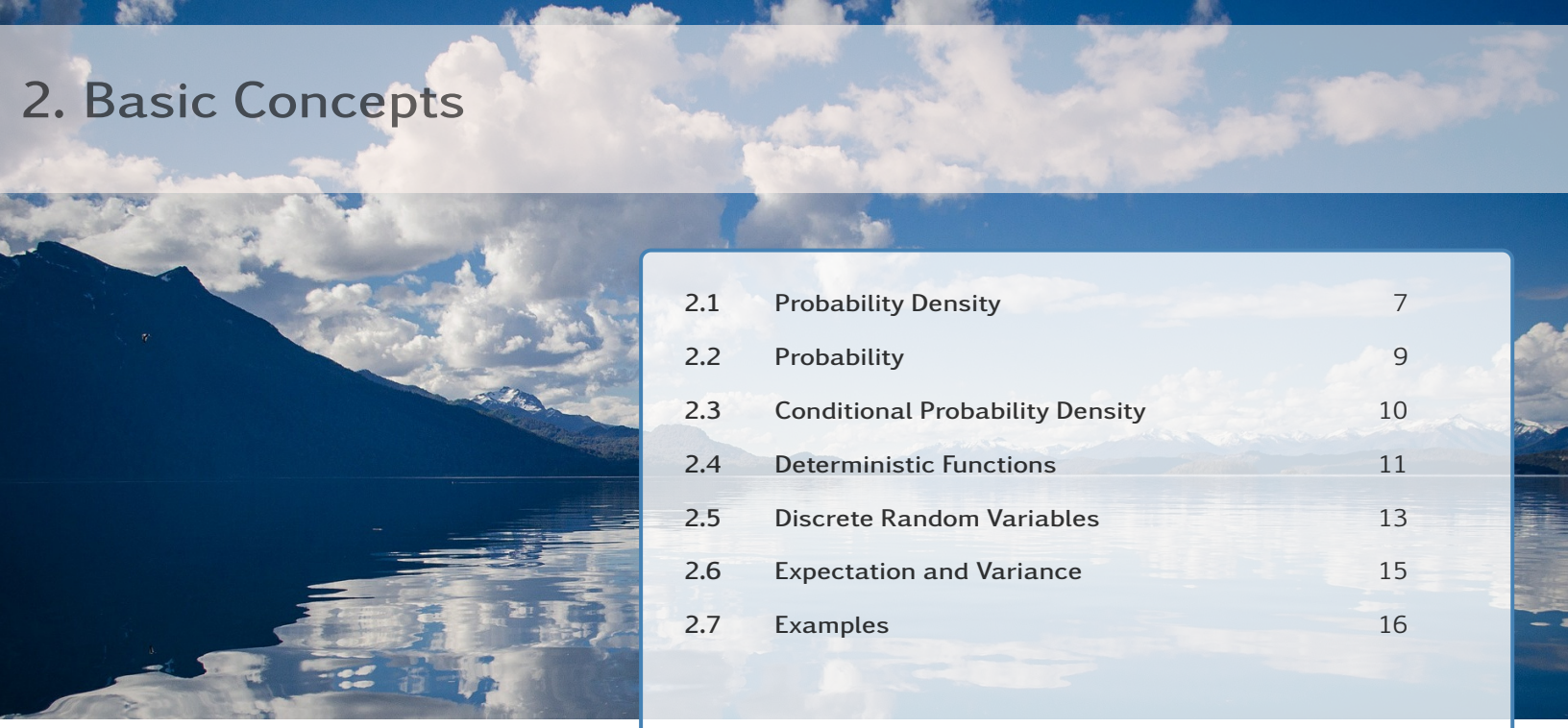
- David Barber's book **Bayesian Reasoning and Machine Learning** [1] is one of the best books, which covers extensively all topics in great detail and clarity.
- Simon Sirca's book **Probability for Physicists** [9] is a good introduction with many examples from physics.
- Glen Cowan's book **Statistical Data Analysis** [3] is a concise and well-written treatise targeting mainly readers of the particle physics community.

Some knowledge in Linear Algebra and Calculus is also required e.g. see Murray Spiegels's book **Schaum's Outline of Advanced Mathematics for Engineers and Scientists** [10] for an excellent coverage of all the mathematical methods and tools used by engineers and scientists. There is also another book by the same author completely dedicated to probability and statistics [11], which consists of 897 fully solved problems and 20 online videos.

There is already a very good book **Introduction to Probability with Mathematica** by Kevin J. Hastings [5] with a similar intention than this one. The second edition is based on Mathematica 7.0 and it is highly recommended as a complementary reading.

Nevertheless the goal of this book is a bit different. We want to cover only the relevant topics and concepts, which are necessary for building and analyzing probabilistic models and applying them to all kind of problems. Therefore it does not claim to be a complete treatise of the subject, but a rather concise introduction to bring the reader as fast as possible into a state to perform computations and the ability to understand its results.

2. Basic Concepts



This first chapter introduces the basic concepts. The starting point in other text books about this topic is typically the definition of probability based on the Kolmogorov axioms. This book chooses a different approach and puts the main focus on the probability density. Everything else including probability is derived from this quantity. The reason for this choice is mainly that the result of almost every non-trivial probabilistic calculation in physics or mathematics is always a probability density. It includes the complete information about a random process and plays a central role in the description of probabilistic systems. As we will see, it contains also deterministic systems as a special case and is therefore a generalization of a function for the case, when uncertainty comes into play.

2.1 Probability Density

The probability density is a generalized function, which describes the distribution of values of an random variable. Generalized means here that it can contain Dirac delta functions. It describes the outcome of an experiment or measurement, where the value of the random variable is determined. The exact value of an outcome is unpredictable, but the frequency of the occurrence of a value is described by the probability density. Even if the probability density is typically a function, the value at a specific point has no direct meaning, only the integral over a certain range of values can be interpreted as probability. The terms probability density function (pdf) or probability distribution is used synonymously for a probability density. So when we write $p(x)$ or $pdf(x)$, it means the probability density associated with the random variable x or x is distributed as $p(x)$. In the function $p(x)$ x is only a placeholder and can be replaced by another symbol e.g. u. In this case we write $p(x = u)$, which means that in the functional form of the pdf associated with the random variable x, we use u as an argument.

The definition of a probability density is given below:

Definition 2.1 (Probability Density) A probability density $p(x)$ is a generalized non-negative function of a variable x with a domain of $[-\infty, +\infty]$ such that

$$\int_{-\infty}^{+\infty} p(x)dx = 1 \tag{2.1}$$

The dimension of a probability density $p(x)$ is the inverse of the dimension of the associated random variable x.

A random variable is defined as

Definition 2.2 (Random Variable) A real-valued continuous variable x is called a random variable if x is distributed as $p(x)$ (also written as $x \sim p(x)$). This means that the distribution of values of a random variable x is completely described by the probability density $p(x)$.

A random variable with a finite domain $[a,b]$ can be described by a probability density, which is zero outside of the interval $[a,b]$.

A probability density of two or more random variables is called a joint probability density e.g. $p(x,y)$ in the case of two random variables x and y. The normalization condition is given by

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} p(x,y)dxdy = 1 \tag{2.2}$$

Integrating only over one random variable e.g. y, the result is again a valid probability density $p(x)$. This operation is called marginalization. Therefore the simplest way to construct a joint (or multi-variate) probability density out of uni-variate probability densities is by multiplication e.g. $p(x,y) = p(x)p(y)$.

One word of caution: if we say that x is distributed as $p(x)$ and y is distributed as $p(y)$, that does not mean that the probability distributions have the same functional form, but it is kept open if they are identical or different. Whenever they are identical, we explicitly mention this and say that x and y are identically distributed.

For multi-variate probability densities we follow the notation of writing the argument in bold letters or with an index range subscript e.g. $p(\mathbf{x})$ or $p(x_{1:n})$ is the short form for the joint probability density $p(x_1, x_2, \ldots, x_n)$.

Example 2.1 (The Univariate Normal Distribution) The most important probability density is the Normal (or Gauss) distribution $\mathcal{N}\left(x|\mu,\sigma^2\right)$, which has the functional form

$$\frac{1}{\sqrt{2\pi\sigma^2}}\exp -\frac{(x-\mu)^2}{2\sigma^2} \tag{2.3}$$

The function has two free parameters μ and σ and fulfills the normalization condition for any value of μ and σ.

$$\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{+\infty}\exp -\frac{(x-\mu)^2}{2\sigma^2}dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\exp -\frac{x^2}{2}dx = 1 \tag{2.4}$$

where the variable transformation $x \to \frac{x-\mu}{\sigma}$ was used.

Example 2.2 (The Multivariate Normal Distribution) The multi-variate Normal (or Gauss) distribution $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\Sigma)$, has the functional form

$$\frac{1}{\sqrt{2\pi\det(\Sigma)}}\exp -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) \tag{2.5}$$

with $\mathbf{x} = x_{1:n}$,$\boldsymbol{\mu} = \mu_{1:n}$ and $\Sigma = \Sigma_{1:n,1:n}$. Σ is called the covariance matrix and is symmetric $\Sigma^T = \Sigma$, so it has only $\frac{n(n+1)}{2}$ independent parameters. Therefore the multi-variate Normal distribution has $\frac{n(n+3)}{2}$ parameters.

An important property of the multi-variate Normal distribution is that the result of a linear transformation $\mathbf{x}' = \mathrm{M}\mathbf{x} + \mathbf{b}$ with a $n \times n$ matrix M and a $n \times 1$ vector $\mathbf{b}$ is again a multi-variate Normal distribution

$$\mathcal{N}(\mathbf{x}'|\boldsymbol{\mu}, \Sigma) = |\det(M)|\,\mathcal{N}\left(\mathbf{x}|\mathrm{M}\boldsymbol{\mu} + \mathbf{b}, \mathrm{M}\Sigma\mathrm{M}^T\right) \tag{2.6}$$

using $\det(\mathrm{M}\Sigma\mathrm{M}^T) = \det(\mathrm{M}^T\mathrm{M}\Sigma) = \det(\mathrm{M})^2 \det(\Sigma)$.

Example 2.3 (The Uniform Distribution) The simplest probability density is the uniform distribution $\mathcal{U}(x|\alpha, \beta)$, which has the functional form

$$\frac{1}{\beta - \alpha}\Theta(\beta - x)\Theta(x - \alpha) \tag{2.7}$$

with $\beta > \alpha$. $\Theta(x)$ is the Heaviside step function, which is defined as

$$\Theta(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geqq 0 \end{cases} \tag{2.8}$$

The two Θ-functions guarantee that the probability density is zero outside the interval $[a, b]$. The normalization condition can be easily verified by

$$\frac{1}{\beta - \alpha}\int_{-\infty}^{+\infty} \Theta(\beta - x)\Theta(x - \alpha)dx = \frac{1}{\beta - \alpha}\int_{\alpha}^{\beta} dx = \frac{\beta - \alpha}{\beta - \alpha} = 1 \tag{2.9}$$

2.2 Probability

Probability is defined in terms of the probability density in the following way.

Definition 2.3 (Probability) The probability of x taking a value between x_1 and x_2 is given by

$$P(x_1 \leqq x \leqq x_2) = \int_{x_1}^{x_2} p(x)dx \tag{2.10}$$

In the case of a joint probability density, the probability of x taking a value between x_1 and x_2 and y taking a value between y_1 and y_2 is given by

$$P(x_1 \leqq x \leqq x_2 \;\&\; y_1 \leqq y \leqq y_2) = \int_{x_1}^{x_2}\int_{y_1}^{y_2} p(x, y)dxdy \tag{2.11}$$

The generalization to the case of a multi-variate probability density with more than 2 random variables is self-explanatory.

The convention to write probabilities with a upper case P and probability density with a lower case p is adapted throughout the book.

We will now show that all the properties of a probability (also known as Kolmogorov axioms) follow from this definition.

Theorem 2.1 (Probability Axioms) The probability satisfies the following 3 axioms:

1. The probability is a non-negative real number.
2. The probability of x taking a value between $[-\infty, +\infty]$ is 1.
3. A set of disjoint intervals $\mathcal{I}_i$ with $i = 1, \ldots, n$ has the probability $\sum_{i=1}^{n} P(\mathcal{I}_i)$.

Proof. Based on the definition of the probability as integral over a probability density, the proof is rather trivial. The first axiom follows directly from the non-negativity of the probability density, the second axiom from the normalization condition and the third axiom from the additivity of integrals over disjoint integration intervals, which simply follows from the interpretation of an integral as an area under a curve. □

Because probability is related to the integral of a probability density, a cumulative probability density is defined as

Definition 2.4 (Cumulative Distribution Function) A cumulative distribution function (cdf) of a random variable x is given by

$$\Phi(x) = \int_{-\infty}^{x} p(x = x')dx' \tag{2.12}$$

The probability of x taking a value between x_1 and x_2 can then be written as the difference of two cumulative distribution functions evaluated at x_2 and x_1.

$$P(x_1 \leqq x \leqq x_2) = \Phi(x_2) - \Phi(x_1) \tag{2.13}$$

2.3 Conditional Probability Density

Almost all probability densities are conditional probability densities, because they depend on various parameters and/or other random variables. We will treat parameters and random variables behind the condition symbol in the same way, which means that we assume that they have a fixed value, when the probability density is evaluated. The only difference is that a random variable can be in front of or behind the condition symbol, whereas parameters can only be placed behind it.

A conditional probability density $p(x|y,\theta)$ means a probability density of x given y (a random variable) and θ (a parameter).

Definition 2.5 (Conditional Probability Density) The conditional probability $p(x|y,\theta)$ is defined by

$$p(x|\ y,\theta) = \frac{p(x,y|\theta)}{p(y|\theta)} \tag{2.14}$$

A similar equation holds for $p(y|x,\theta)$

$$p(y|\ x,\theta) = \frac{p(x,y|\theta)}{p(x|\theta)} \tag{2.15}$$

Eliminating the joint probability density $p(x,y|\theta)$ leads to the Bayes rule

$$p(x|\ y,\theta) = \frac{p(y|\ x,\theta)p(x|\theta)}{p(y|\theta)} \tag{2.16}$$

A joint probability density $p(\mathbf{x}|\mathbf{y},\boldsymbol{\theta})$ can be decomposed into univariate conditional probability densities by iteratively applying Bayes rule.

$$p(\mathbf{x}|\mathbf{y},\boldsymbol{\theta}) = p(x_{1:n}|\mathbf{y},\boldsymbol{\theta}) = p(x_1|x_{2:n},\mathbf{y},\boldsymbol{\theta})\underbrace{p(x_2|x_{3:n},\mathbf{y},\boldsymbol{\theta})\overbrace{\ldots p(x_{n-1}|x_n,\mathbf{y},\boldsymbol{\theta})p(x_n|\mathbf{y},\boldsymbol{\theta})}^{p(x_{3:n}|\mathbf{y},\boldsymbol{\theta})}}_{p(x_{2:n}|\mathbf{y},\boldsymbol{\theta})} \tag{2.17}$$

This decomposition is not unique. Any permutations of $\mathbf{x}$ is possible here.

The question is: what is the best choice for this decomposition?

There is no general answer here, but if you are able to exploit the conditional independence of random variables as often as possible by using

$$p(x|y,\theta) = p(x|\theta) \tag{2.18}$$

then the individual terms in the product will simplify significantly. The statement x is conditionally independent from y means that the functional form of the probability density of x is the same regardless of the value of y.

Expressions with a product of probability densities will occur very often and are subject to numerical underflow. It is therefore recommended to perform numerical evaluations always with the logarithm of a probability density. It does not matter if it is the natural logarithm or the logarithm with base 10. The crucial point is that a product of pdf's is converted into a sum of logarithms of pdf's.

2.4 Deterministic Functions

We will now show that every function can be written as a conditional probability density. This will allow us to treat everything as probability density and mix deterministic and random behavior in our models.

Let's study the Normal distribution in the limit $\sigma^2 \to 0$, where the uncertainty of observing the value μ changes to certainty and the outcome of the random variable x is always μ.

Example 2.4 (The Normal Distribution with zero variance)

$$\lim_{\sigma^2 \to 0} \mathcal{N}\left(x \,|\, \mu, \sigma^2\right) = \lim_{\sigma^2 \to 0} \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(x-\mu)^2}{2\sigma^2} = \delta(x-\mu) \tag{2.19}$$

As we can see from the example, the probability density becomes a Dirac delta function with the functional relation $x = \mu$ as argument.

Definition 2.6 (Deterministic Probability Density) Any function $y = f(x)$ can be written as conditional probability density in the following way.

$$p(y \,|\, x) = \delta(y - f(x)) \tag{2.20}$$

A common use-case is that we have a conditional probability density, which will be become deterministic by including hidden (or latent) random variables. The random character of a quantity could be completely originate from a another random variable and as soon as this other random variable has been identified and included, the conditional probability density can be replaced by an expression of the form shown above if the functional dependency on the hidden variable is known.

Another consequence is that an integration over y can be carried out immediately using

$$\int_{-\infty}^{+\infty} \delta(y - f(x))\, g(y) dy = g(f(x)) \tag{2.21}$$

Another use-case are variable transformations from a random variable x with a known probability density $p(x)$ to another variable $y = f(x)$.

The question is: what is the pdf of y?

The starting point for the case with two random variables x and y is always the joint probability density $p(x, y)$. Because we are only interested in the pdf of y, we marginalize over x.

$$p(y) = \int_{-\infty}^{+\infty} p(x, y) dx = \int_{-\infty}^{+\infty} p(y \,|\, x) p(x) dx = \int_{-\infty}^{+\infty} \delta(y - f(x))\, p(x) dx \tag{2.22}$$

Using the following identity for the Dirac delta function

$$\delta(y - f(x)) = \sum_{i=1}^{n} \frac{1}{|f'(x_i)|} \delta(x - x_i(y)) \tag{2.23}$$

where $x_i(y)$ are all n solutions of $y = f(x)$ and $f'(x_i)$ is the derivative of $f(x)$ with respect to x evaluated at $x = x_i(y)$, one can carry out the integration over x with the result

$$p(y) = \sum_{i=1}^{n} \frac{1}{|f'(x_i)|} p\left(x = x_i(y)\right) \tag{2.24}$$

Example 2.5 (The Square of a Random Variable) What is the pdf of $y = x^2$ if $x \sim p(x)$? The equation $y = x^2$ has two solutions $x_1 = +\sqrt{y}, x_2 = -\sqrt{y}$. The derivative of x^2 is $2x$.

$$p(y) = \sum_{i=1}^{2} \frac{1}{|2x_i|} p\left(x = x_i(y)\right) = \frac{1}{2\sqrt{y}} \left(p\left(x = +\sqrt{y}\right) + p\left(x = -\sqrt{y}\right)\right) \tag{2.25}$$

y is only defined for values in the range $[0, \infty]$. Therefore we have to add $\Theta(y)$.

$$p(y) = \frac{1}{2\sqrt{y}} \left(p\left(x = +\sqrt{y}\right) + p\left(x = -\sqrt{y}\right)\right) \Theta(y) \tag{2.26}$$

Example 2.6 (The Square of a Gaussian Zero-Mean Random Variable) What is the pdf of $y = x^2$ if $x \sim \mathcal{N}(x|0, \sigma^2)$? The result from the previous example can be used to obtain

$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2 y}} \exp\left(-\frac{y}{2\sigma^2}\right) \Theta(y) \tag{2.27}$$

The probability density is infinite at $y = 0$, but it is still a valid distribution, because the normalization condition is fulfilled and it is non-negative for all values of y. A probability density can be singular at some points, as long as the integral over it is finite.

2.5 Discrete Random Variables

So far, we have studied continuous random variables and their associated probability densities. Next, we look at discrete random variables, which only take values from a discrete countable set $x_1, x_2, \ldots, x_n$.

How can we construct a probability density for discrete random variables?

We assign a probability P_i to each x_i for $i = 1, \ldots, n$ and the corresponding discrete probability density is given by

Definition 2.7 (Discrete Probability Density)

$$p(x) = \sum_{i=1}^{n} P_i \delta(x - x_i) \tag{2.28}$$

Note that x is still a continuous variable, but due to the Dirac delta functions it is restricted to the discrete values x_i. We write here the coefficients with a capital P to indicate that these numbers are indeed probabilities. The normalization condition simplifies to a sum over all probabilities.

$$\int_{-\infty}^{+\infty} p(x) dx = \sum_{i=1}^{n} P_i = 1 \tag{2.29}$$

The definition shows that in the case of discrete random variables, probability densities $p(x)$ reduce to probabilities $P(x)$ and integrals to sums over all the possible values of x.

Example 2.7 (The Binomial Distribution) The binomial distribution is one of the most important discrete probability densities. x_k is e.g. the outcome of a sequence of n coin flips with k heads and $n-k$ tails. The probability of head in a coin flip is p, while the probability of tail is $1-p$.

$$\mathcal{B}(x|\ n,p) = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k} \delta(x-x_k) \tag{2.30}$$

Let' say that we flip a fair coin with $p = \frac{1}{2}$ $n = 3$ times and we want to calculate the probability that we have $k = 2$ heads and $n-k = 1$ tails. All coin flip sequences of the form hht, hth, thh contribute to it. Therefore the probability is $\frac{3}{8}$. The same probability is obtained in the case $k = 1$ heads and $n-k = 2$ tails. For the case $k = 3$ heads and 0 tails there is only one possibility: hhh. Same for $k = 0$ heads and $n-k = 3$ tails. Therefore the probability in these cases is $\frac{1}{8}$. The sum of all probabilities is $\frac{1+3+3+1}{8} = 1$.

Example 2.8 (The Poisson Distribution) In the limit $n \to \infty$ and $\lambda = np$ remains finite, the binomial distribution becomes the Poisson distribution given by

$$\mathcal{P}(x|\lambda) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \exp(-\lambda) \delta(x-x_k) \tag{2.31}$$

This distribution describes the probability of observing k events given a mean event count of λ.

Theorem 2.2 (Sum of two Poisson distributions) Show that $z = x+y$ is distributed as a Poisson-distribution $P(z|\lambda_1+\lambda_2)$ if $y \sim \mathcal{P}(x|\lambda_1)$ and $y \sim \mathcal{P}(y|\lambda_2)$.

Proof.

$$p(z) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(z,x,y)dxdy = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(z|x,y)\mathcal{P}(x|\lambda_1)\mathcal{P}(y|\lambda_2)dxdy \tag{2.32}$$

Bayes rule was used in last step. $p(z|x,y)$ is a deterministic function $z = x+y$ and is given by $\delta(z-x-y)$. The y-integration can be carried out and gives the following result.

$$p(z) = \int_{-\infty}^{+\infty} P(x|\lambda_1)P(z-x|\lambda_2)dx \tag{2.33}$$

Inserting the expressions for the Poisson distribution and carrying out the x-integration leads to

$$p(z) = \exp(-\lambda_1-\lambda_2) \sum_{k=0}^{\infty} \frac{\lambda_1^k}{k!} \sum_{l=0}^{\infty} \frac{\lambda_2^l}{l!} \delta(z-x_k-y_l) \tag{2.34}$$

We set now $z_n = x_k + y_l$ with $n = k+l$ by introducing a third sum $\sum_{n=0}^{\infty} \delta_{n,k+l} = 1$ and carrying

out the summation over l

$$p(z) = \exp(-\lambda_1 - \lambda_2) \sum_{n=0}^{\infty} \frac{1}{n!} \underbrace{\sum_{k=0}^{n} n! \frac{\lambda_1^k}{k!} \frac{\lambda_2^{n-k}}{(n-k)!}}_{(\lambda_1+\lambda_2)^n} \delta(z - z_n) \tag{2.35}$$

We used here $\delta_{n,k+l} = \delta_{n-k,l}$, which replaces l by $n-k$ and limits the sum over k to n, because l is positive and $n-k$ is negative for $k > n$ and the Kronecker delta is therefore always zero. □

Exercise 2.1 Show that $z_n = \sum_{i=1}^{n} x_i$ is distributed as a Poisson-distribution $P\left(z_n | \sum_{i=0}^{n} \lambda_i\right)$ if $x_i \sim \mathcal{P}(x_i | \lambda_i)$. Proof this statement by induction using $z_n = z_{n-1} + x_n$.

Exercise 2.2 Show that the probability density of $z = x + y$ can be written as the convolution of the probability densities of x and y.

$$p(z) = \int_{-\infty}^{+\infty} p(x)\, p(y = z - x)\, dx \tag{2.36}$$

Exercise 2.3 Show that $z = x + y$ is distributed as $\mathcal{N}\left(z | \mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2\right)$ if $y \sim \mathcal{N}(x | \mu_1, \sigma_1^2)$ and $y \sim \mathcal{N}(y | \mu_2, \sigma_2^2)$. The first part of the derivation is similar to the one for the Poisson distribution with the result

$$p(z) = \int_{-\infty}^{+\infty} \mathcal{N}\left(x | \mu_1, \sigma_1^2\right) \mathcal{N}\left(z - x | \mu_2, \sigma_2^2\right) dx \tag{2.37}$$

The x integration can be simplified using the variable transformation $x \to \frac{x - \mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}$.

2.6 Expectation and Variance

While the probability density contains all the information needed to perform calculations, it is often useful to characterize its shape by two values: expectation and variance.

Definition 2.8 (Expectation)

$$\mu = E[x] = \int_{-\infty}^{+\infty} x p(x) dx \tag{2.38}$$

Definition 2.9 (Variance)

$$\sigma^2 = V[x] = E[(x - \mu)^2] = \int_{-\infty}^{+\infty} (x - \mu)^2 p(x) dx \tag{2.39}$$

The normal distribution $\mathcal{N}\left(x | \mu, \sigma^2\right)$ is completely characterized by these values, while other distributions have non-zero expectation value of higher powers of $x - \mu$.

Exercise 2.4 Show that the expectation and variance of a random variable x distributed as a Poisson distribution $\mathcal{P}(x|\lambda)$ is given by $E[x] = \lambda$ and $V[x] = \lambda$ if the possible values x can take are $x_k = k$.

Exercise 2.5 Show that the expectation and variance of a random variable x distributed as a uniform distribution $\mathcal{U}(x|\alpha,\beta)$ is given by $E[x] = \frac{\alpha+\beta}{2}$ and $V[x] = \frac{(\beta-\alpha)^2}{12}$.

2.7 Examples

Example 2.9 (The Monty Hall Problem) This is a famous and non-trivial example of a probabilistic problem, where intuition will give you the wrong result. The problem statement is as follows:

Monty Hall, after which the problem is named, was a moderator of a game show and a participant has to choose between 3 doors. Behind one door a prize is hidden. After the choice has been made e.g. door 2, the moderator opens a door with no prize e.g. door 1, and asks the participant if he wants to choose the remaining door (in this case door 3) or if he wants to stay with his decision (door 2).

The question is now: What is the probability to win if the participant stays with his first decision or if he chooses the remaining door?

The difficult part is to cast the problem into a manageable form, which means, selecting proper random variables to make the computation as easy as possible. A discrete random variable x_1 can be associated with the action that the participant chooses a door randomly at the beginning and the choice is either a door with a prize behind it $x_1 = p$ or a door with no prize $x_1 = n$ (*note that this is not known, when the choice happens, but the crucial point is that the probability for its occurrence can be calculated easily with the result* $P(x_1 = p) = \frac{1}{3}$ *and* $P(x_1 = n) = \frac{2}{3}$, *because there are only 3 doors and one contains a prize and the two others not*).

Another discrete random variable x_2 will be associated with the outcome of the game (participant wins $x_2 = w$ or looses $x_2 = l$). This is the random variable, which is needed to answer the question above.

Furthermore we introduce a parameter θ, which describes the two possible ways: participant stays with his first decision ($\theta = 0$) or participant chooses the remaining door ($\theta = 1$).

After the random variables and parameters have been defined, the starting point is always the joint probability $P(x_2, x_1|\theta)$. Because we are only interested in the probability of winning the game, we can marginalize over x_1 and use Bayes rule to separate $P(x_1)$,

which is known.

$$P(x_2 = w|\theta) = \sum_{x_1=p,n} P(x_2 = w, x_1|\theta) = \sum_{x_1=p,n} P(x_2 = w|x_1, \theta)P(x_1) \tag{2.40}$$

We only need to calculate the conditional probability $P(x_2|x_1, \theta)$. We will see that this quantity is completely deterministic.

Let's study first the case $\theta = 0$.

In this case if x_1 was already the right door with the prize, the participant will win, and if x_1 was the wrong door, he will loose regardless if the moderator opens another door without a prize.

$$P(x_2|x_1, \theta = 0) = \begin{cases} 1 & x_2 = w \ \& \ x_1 = p \\ 0 & x_2 = l \ \& \ x_1 = p \\ 1 & x_2 = l \ \& \ x_1 = n \\ 0 & x_2 = w \ \& \ x_1 = n \end{cases} \tag{2.41}$$

Next we handle the case $\theta = 1$.

In this case if x_1 was already the right door with the prize, the participant will loose when changing to the remaining door and if x_1 was the wrong door, he will win, because the moderator has opened the door without no prize, so the remaining door must contain the prize.

$$P(x_2|x_1, \theta = 1) = \begin{cases} 0 & x_2 = w \ \& \ x_1 = p \\ 1 & x_2 = l \ \& \ x_1 = p \\ 0 & x_2 = l \ \& \ x_1 = n \\ 1 & x_2 = w \ \& \ x_1 = n \end{cases} \tag{2.42}$$

So the final result is

$$P(x_2 = w|\theta) = \begin{cases} 1 \cdot \frac{1}{3} + 0 \cdot \frac{2}{3} = \frac{1}{3} & \theta = 0 \\ 0 \cdot \frac{1}{3} + 1 \cdot \frac{2}{3} = \frac{2}{3} & \theta = 1 \end{cases} \tag{2.43}$$

The remarkable result is that making a new choice increases your chances to win considerably. Naively one would have thought that the probability in this case is $\frac{1}{2}$, because after the moderator has opened a door without the prize, two doors are remaining and there is an equal chance to get the right door with the prize. But unfortunately this intuitively convincing argument is wrong.

The following code shows a Monte Carlo simulation with the following variables and functions:
`doorsWithPrize` is a randomly generated list of door numbers, which contain a prize,
`firstChoice` is a randomly generated list of door numbers from the participant's first choice (this variable will be used to evaluate the case participant stays with his first choice),

`doorsWithoutPrize` is a randomly generated list of the doors opened by the moderator after the first choice was made,
`secondChoice` is a list of door numbers from the participant's second choice (this variable will be used to evaluate the case participant chooses the remaining door),
`ProbabilityWinningPrize` is a function to calculate the probability to win a prize for a given choice (it simply counts the number of elements in the choice list, which agree with the corresponding item in the `doorsWithPrize` list).

```
doorsWithPrize = RandomChoice[{1, 2, 3}, 1000];
firstChoice = RandomChoice[{1, 2, 3}, Length[doorsWithPrize]];
doorsWithoutPrize = MapThread[
                        RandomChoice[Complement[{1, 2, 3}, {#1, #2}]] &,
                        {doorsWithPrize, firstChoice}];
secondChoice = MapThread[
                        RandomChoice[Complement[{1, 2, 3}, {#1, #2}]] &,
                        {doorsWithoutPrize, firstChoice}];
ProbabilityWinningPrize[choice_] := Apply[Plus,
                                          MapThread[Boole[#1 == #2] &,
                                          {doorsWithPrize, choice}]
                                          ] / Length[doorsWithPrize] // N;
Print["Probability of winning by staying with your first choice = ",
      ProbabilityWinningPrize[firstChoice]];
Print["Probability of winning by choosing the remaining door = ",
      ProbabilityWinningPrize[secondChoice]];
```

The probability of winning with 1000 samples is `0.32` in the case of staying with the first choice and `0.68` in the case of choosing the remaining door. Using 1000000 samples the values are `0.333` and `0.667`, respectively.

3. Density Estimation

This chapter deals with the problem how a probability density can be estimated from data.

3.1 Histograms

A histogram divides the domain of a random variable x into disjunct bins and counts the number of outcomes of x in each bin. If $x \sim p(x)$, the normalized count in bin $[x_i, x_{i+1}]$ is given by

$$P_i = P\left(x_i \leqq x \leqq x_{i+1}\right) = \int_{x_i}^{x_{i+1}} p(x)dx = \lim_{N\to\infty} \frac{N_i}{N} \tag{3.1}$$

The count N_i can only take integer values, therefore $N_i = [P_i N]$, where the symbol $[x]$ means nearest integer of x and N is the total number of outcomes.

Let's assume that we have collected data about the random variable x as $\mathcal{D} = \{x_{1:N}\}$. We assign each data point x_k to a bin and count the number of data points in each bin. If the bins do not cover the whole range of x, there will be data points, which can not be assigned to a bin. These data points are called outliers.

If we represent the bins in terms of a center point $\tilde{x}_i = \frac{x_i + x_{i+1}}{2}$ and the width $w_i = \frac{x_{i+1} - x_i}{2}$, we can write the bin b_i as $[\tilde{x}_i - \frac{w_i}{2}, \tilde{x}_i + \frac{w_i}{2}]$.

Then the N_i can be expressed as

$$N_i = \sum_{k=1}^{N} \mathcal{I}\left(x_k - \tilde{x}_i; w_i\right) \tag{3.2}$$

where $\mathcal{I}$ is the indicator function defined as

$$\mathcal{I}(x; w) = \begin{cases} 1 & x \in [-\frac{w}{2}, +\frac{w}{2}] \\ 0 & otherwise \end{cases} \tag{3.3}$$

Then the probability density estimate $\hat{p}(x)$ can be written as

$$\hat{p}(x) = \frac{1}{N} \sum_{i=1}^{b} \frac{N_i}{w_i} \mathcal{I}\left(x_k - \tilde{x}_i; w_i\right) \tag{3.4}$$

where b is the number of bins.

There are some best practice rules for choosing the number of bins b and the bin width w:

$$b = [1 + \log_2(N)] \tag{3.5}$$

$$w = 3.5sN^{-\frac{1}{3}} \tag{3.6}$$

where s is the standard deviation of the data sample.

3.2 Maximum Likelihood

If the physical process how data is generated is understood, the functional form of the probability density $p(\mathbf{x}|\boldsymbol{\theta})$ is typically known. In these cases the goal is to estimate the parameters $\boldsymbol{\theta}$ of a probability density from data.

Definition 3.1 (Likelihood) The likelihood is defined by

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{N} p(x_i|\boldsymbol{\theta}) \tag{3.7}$$

where $x_i \in \mathcal{D} = \{x_{1:N}\}$. It is a function of the parameter of the probability density, not a distribution.

Using the logarithm of the likelihood is better from the view point of numerical stability.

Definition 3.2 (Loglikelihood) The loglikelihood is defined by

$$\log(\mathcal{L}(\boldsymbol{\theta})) = \sum_{i=1}^{N} \log(p(x_i|\boldsymbol{\theta})) \tag{3.8}$$

An optimal estimate of the parameters $\hat{\boldsymbol{\theta}}$ can be found by solving $\frac{\partial \log(\mathcal{L}(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} = 0$.

3.3 Sampling

In the previous sections we studied the reconstruction of a probability density from data. Here we investigate the inverse problem: the generation of data from a given probability density. This process is called sampling.

It is usually quite easy to generate random numbers r from a uniform distribution $\mathcal{U}(r|\alpha = 0, \beta = 1)$. Almost every programming language has a function for it. The goal is now to transform the sequence $r_{1:N}$ to a sequence $x_{1:}$ so that x is distributed as $p(x)$.

One way to do this is to find a transformation function $x(r)$ for $0 \leqq r \leqq 1$.

$$\Phi(x(r)) = \int_{-\infty}^{x(r)} p(x)dx = \int_{-\infty}^{r} \mathcal{U}(r = r'|\alpha = 0, \beta = 1)dr' = r \tag{3.9}$$

The lhs is the cdf evaluated at $x(r)$. Therefore the function $x(r)$ is the inverse of the cdf evaluated at r.

$$x(r) = \Phi^{-1}(r) \tag{3.10}$$

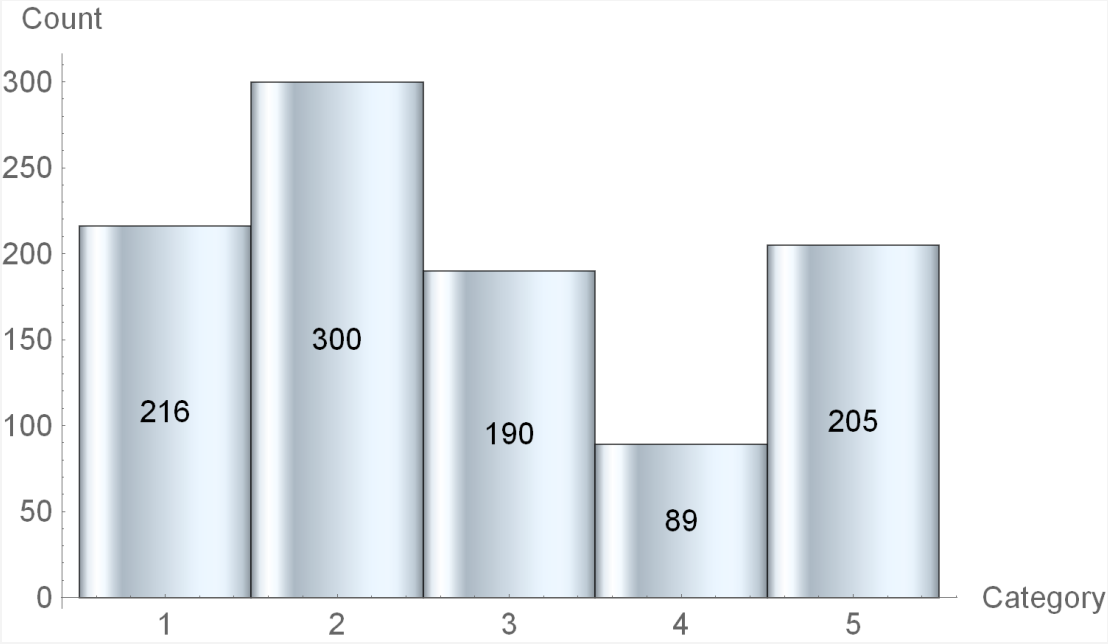
Example 3.1 (Sampling from a Discrete Distribution) Let's assume that we have an array of probabilities $P_{1:N}$ with $\sum_{i=1}^{N} P_i = 1$ and we want to sample an index variable $j \in [1, 2, \ldots, N]$ from this distribution.

In this case the cumulative distribution is given by $\Phi(j) = \sum_{i=1}^{j} P_i$. We iterate the index j from N down to 1 and stop as as soon as the condition $\Phi(j) \leqq r$ is fulfilled. This procedure generates an index sequence with the desired distribution.

The Wolfram Language provides the function `EmpiricalDistribution` to construct a pdf from a list of weights, which need not to be normalized. Using `RandomVariate` 1000 samples are drawn from this pdf and are used to fill a histogram.

```
weights = {200, 300, 200, 100, 200};
categories = Range[1, Length[weights]];
dist = EmpiricalDistribution[weights -> categories];
data = RandomVariate[dist, 1000];
Histogram[data,
          LabelingFunction -> Center,
          ChartElementFunction -> "GlassRectangle",
          ChartStyle -> LightBlue,
          AxesLabel -> {"Category", "Count"}]
```

The histogram shows the counts of the data samples generated from an empirical distribution for each category. The counts are indeed numerically close to the given weights.



Example 3.2 (Normal-Distributed Random Numbers) The trick here is to perform the transformation on the joint pdf of x and y, where both random variables are identically distributed according to $\mathcal{N}(\_\,|\,\mu = 0, \sigma^2 = 1)$.

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \mathcal{N}(x\,|\,\mu = 0, \sigma^2 = 1)\mathcal{N}(y\,|\,\mu = 0, \sigma^2 = 1)dxdy = 1 \tag{3.11}$$

Using the transformation $x = \rho\cos(\phi)$ and $y = \rho\sin(\phi)$, the double integral can written as

$$\frac{1}{2\pi}\int_0^{2\pi} d\phi \int_0^{+\infty} \exp\left(-\frac{\rho^2}{2}\right)\rho d\rho = \int_0^1 d\phi' \int_0^{+\infty} \exp\left(-\rho'\right) d\rho' = 1 \tag{3.12}$$

where the new variables ρ' and ϕ' are given by $\rho' = \frac{\rho^2}{2}$ and $\phi' = \frac{\phi}{2\pi}$. So ϕ' is uniformly distributed in the range $[0,1]$ and ρ' is exponentially distributed with $\xi = 1$.

We leave it as an exercise to show that $\rho'(r) = \ln(r)$ and therefore $\rho(r) = \sqrt{-2\ln(r)}$.

The transformation function for ϕ is given by $\phi(r) = 2\pi r$. Because ϕ and ρ are independent random variables, we must also use independent random number sequences r and s to evaluate them.

Therefore the transformation for x is given by

$$x(r,s) = \sqrt{-2\ln(r)}\cos(2\pi s) \tag{3.13}$$

and for y by

$$y(r,s) = \sqrt{-2\ln(r)}\sin(2\pi s) \tag{3.14}$$

The algorithm is called Box-Muller method [2].

Exercise 3.1 Show that the function $x(r)$ for the exponential distribution $p(x) = \frac{1}{\xi}\exp\left(-\frac{x}{\xi}\right)$ is given by $x(r) = -\xi\ln(1-r)$ or $x(r) = -\xi\ln(r)$.

Exercise 3.2 Show that how the method above has to be extended to generate random numbers distributed as $\mathcal{N}(\_|\mu,\sigma^2)$.

Finding a function $x(r)$ analytically is not always possible. Another approach, which is applicable to a wider range of distributions, is rejection sampling [8].

Definition 3.3 (Rejection Sampling) The algorithm works in the following way:

- Choose a number M and a distribution $p(y)$ in such a way that the condition $p(x = u) \leqq Mp(y = u)$ holds everywhere on the domain of x and drawing samples from $p(y)$ is known.
- Draw samples y_k from a distribution $p(y)$ and r_k from a uniform distribution $\mathcal{U}(r|\alpha = 0, \beta = 1)$ for $k = 1 : N$.
- Check for all k the condition $r_k < \frac{p(x=y_k)}{Mp(y=y_k)}$. If it is true, accept y_k as a sample from $p(x)$, otherwise reject the sample.

If $p(x)$ is bounded to a finite interval $[a,b]$, the algorithm can be simplified:

- Choose p_{max} as the maximum value of $p(x)$ for $x \in [a,b]$
- Draw samples $x_k = (b-a)s_k + a$ from a uniform distribution $\mathcal{U}(s|\alpha = 0, \beta = 1)$ and r_k from a uniform distribution $\mathcal{U}(r|\alpha = 0, \beta = 1)$ for $k = 1 : N$.

- Check for all k the condition $r_k < \frac{p(x=x_k)}{p_{max}}$. If it is true, accept x_k as a sample from $p(x)$, otherwise reject the sample.

The rejection rate determines the efficiency of the algorithm.

Finding a pdf, which is easy to sample from and has a similar functional form as the distribution of interest, can be achieved with the help of a Gaussian Mixture Model.

Definition 3.4 (Gaussian Mixture Model) The Gaussian Mixture Model (GMM) is defined by

$$p_{GMM}(x|\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \sum_{i=1}^{N} \alpha_i \mathcal{N}\left(x|\mu_i, \sigma_i^2\right) \tag{3.15}$$

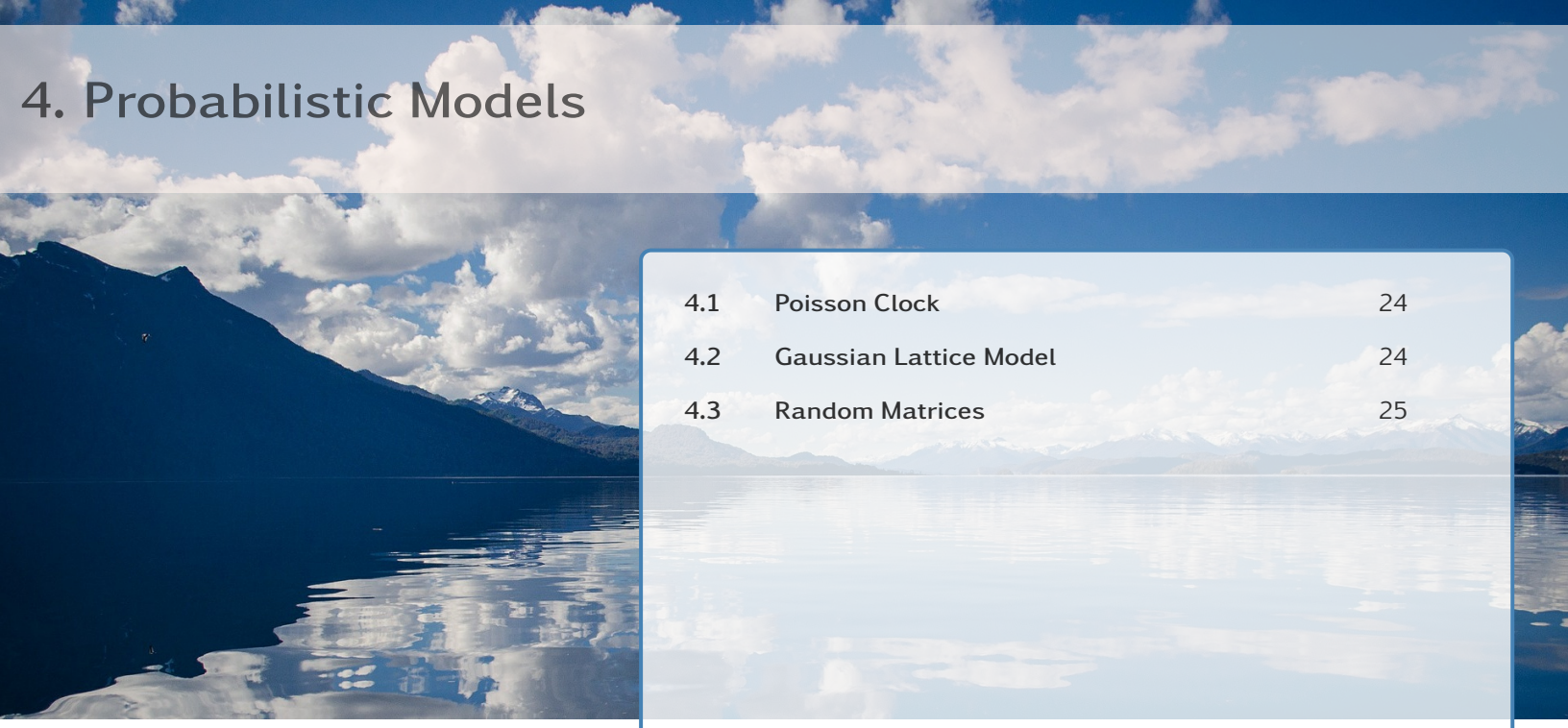
with $\sum_{i=1}^{N} \alpha_i = 1$

The parameters $,\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2$ determines the shape of the distribution and can be chosen to approximate a given pdf $p(x)$ as closely as possible.

Sampling from an Gaussian Mixture Model is straightforward:

- Sample an index j from the empirical distribution of $\alpha_{1:N}$
- Sample x from the normal distribution $\mathcal{N}\left(x|\mu_j, \sigma_j^2\right)$

4. Probabilistic Models



This chapter covers the whole range of probabilistic models from simple to complex ones. We show that the simple models can be used as building blocks for simulating real-world systems.

4.1 Poisson Clock

A very simplified model of a clock is the Poisson clock, where the tick counts x are Poisson-distributed.

Definition 4.1 (Poisson Clock)

$$\mathcal{P}(x|\lambda \to \nu t) = \sum_{k=0}^{\infty} \frac{\nu^k t^k}{k!} \exp(-\nu t)\delta(x-k) \tag{4.1}$$

ν is the clock rate (number of ticks in a certain time interval) and t is the elapsed time. Because the expectation value is $\mu = \lambda t$, the tick count is a measure of time.

However, this is not a good model of a real clock, because there is only one parameter ν, which controls the clock rate, and the clock noise is completely determined by this choice. In a real clock model you can control the rate and the noise independently.

Nevertheless, the Poisson clock is used very often for e.g. simulating the arrival of costumers in a shop or the number of decays of a radioactive material. It is also a building block of the Ptolemy II simulation framework [4].

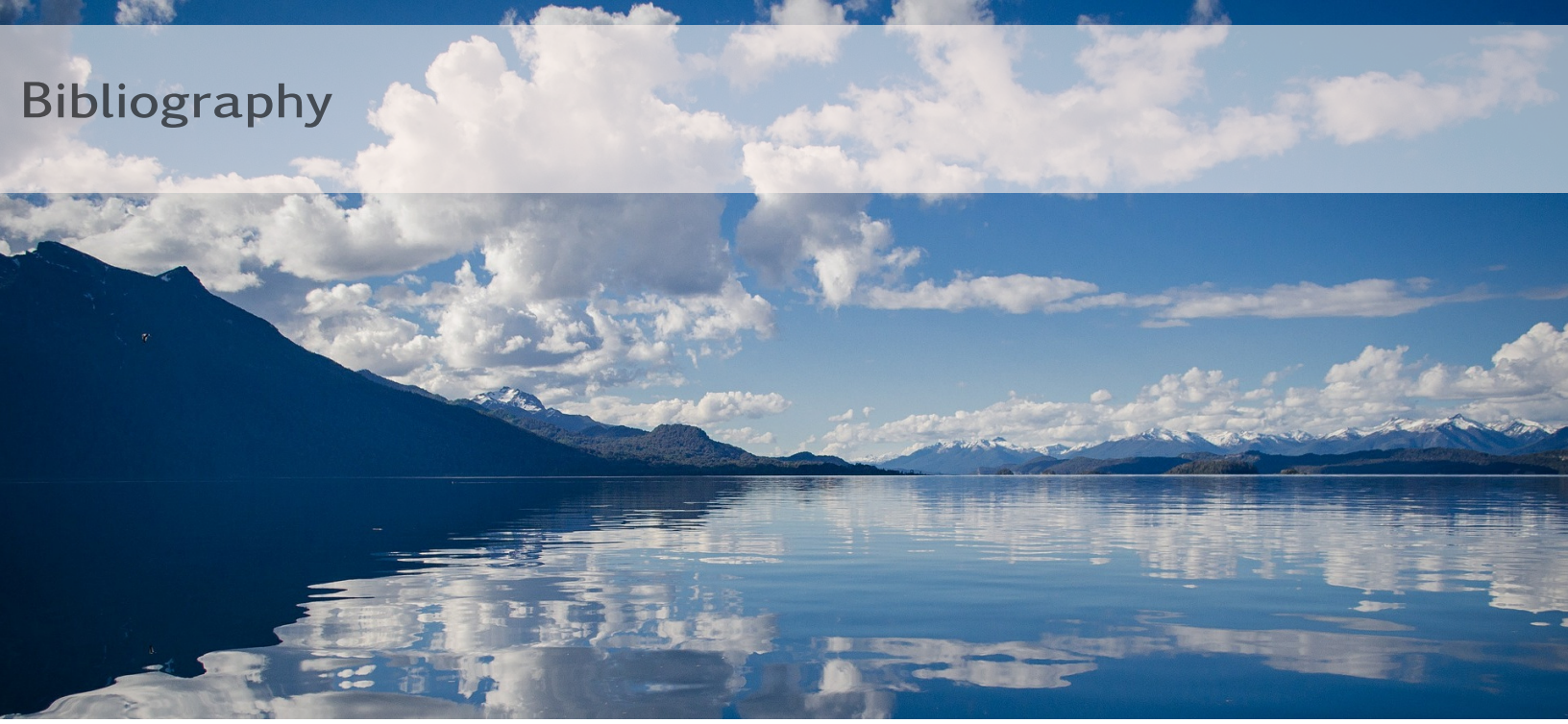
4.2 Gaussian Lattice Model

The Gaussian Lattice Model is a variant of the famous Ising Model [7]. We want to study this model in the one-dimensional case, where it can be solved exactly.

4.3 Random Matrices

A random matrix m is $N \times N$ matrix with its elements $m_{ij} \sim p(x)$ [6].

Bibliography



[1] David Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2011. URL: http://www.cs.ucl.ac.uk/staff/d.barber/brml.

[2] George Edward Pelham Box and Mervin Edgar Muller. "A note on the generation of random normal deviates". In: *Annals of Mathematical Statistics* 29.2 (1958), pp. 610–611.

[3] Glen Cowan. *Statistical data analysis*. Oxford University Press, USA, 1998.

[4] Johan Eker et al. "Taming heterogeneity - the Ptolemy approach". In: *Proceedings of the IEEE* 91.1 (Jan. 2003), pp. 127–144.

[5] Kevin J. Hastings. *Introduction to Probability with Mathematica*. CRC Press, 2001.

[6] Giacomo Livan, Marcel Novaes, and Pierpaolo Vivo. *Introduction to Random Matrices - Theory and Practice*. Springer Briefs in Mathematical Physics 26. Springer, 2017.

[7] Giuseppe Mussardo. *Statistical Field Theory. An Introduction to Exactly Solved Models of Statistical Physics*. Oxford University Press, 2010.

[8] John von Neumann. "Various techniques used in connection with random digits. Monte Carlo methods". In: *Nat. Bureau Standards* 12 (1951), pp. 36–38.

[9] Simon Sirca. *Probability for physicists*. Graduate texts in physics. Berlin: Springer, 2016.

[10] Murray Spiegel. *Schaum's Outline of Advanced Mathematics for Engineers and Scientists*. Schaum's Outline Series, 2009.

[11] Murray Spiegel. *Schaum's Outline of Probability and Statistics*. Schaum's Outline Series, 2012.

Index