Chapter 1

Probabilistic modeling and inference

By the end of this chapter, we will have presented

- *Data oriented modeling*
- *Random variables and their properties*
- *An overview of inverse problem solving*

1.1 Modeling with data

If experimental observations or, put concretely, binaries on a screen were all we ever cared about, then no experiment would require modeling or interpretation and the remainder of this book would be unnecessary. But binaries on a screen do not constitute knowledge. They constitute *data*. Put differently, Quantum Mechanics like any scientific knowledge is not self-evident from the pixelated outcome on a camera chip of a modern incarnation of a Young's two-slit interference experiment.

In the Natural Sciences, *models* of physical systems provide mathematical frameworks from which we unify disparate pieces of information. These include conceptual notions such as symmetries, fundamental constituents and other postulates as well as scientific *measurements* and, even more generally, empirical observations of any form. If we think of direct observations as data in particular, at least for now, we can think of mathematical models as a way of compressing or summarizing these data.

Data summaries may be used to make predictions about physical conditions we may encounter in the future, such as in new experiments, or to interpret and describe an underlying physical system already probed in past experiments. For example, with time-ordered data we may be interested in learning equations of motion or kinetic schemes. Or, already knowing a kinetic scheme sufficiently well from past experiments or fundamental postulates, we may only be interested in learning the noise characteristics of a new piece of equipment on which future experiments will be run. Thus, models may be aimed at discovering new Science as well as at devising careful controls to get a better handle on error bars and, more broadly, even at designing new experiments altogether.

1.1.1 Why do we obtain models from raw data?

Experimental data rarely provide direct insight on the physical conditions and systems of interest. At the very least, measurements are *corrupted* by unavoidable noise and, as a result, models obtained from experimental data are unavoidably probabilistic. So, we ask: *how should we, the scientific community, go about obtaining models from imperfect data?*

Note 1.1: Obtaining models from data

Data can be time and labor intensive to acquire. Perhaps more importantly, every datum in a dataset encodes information. In light of this, we re-pitch our question and ask: *how should we go about obtaining models* efficiently *and* without compromising *the information encoded in the data?*

The key is to start from data acquired in experiments and arrive at models with a minimal amount of data pre-processing, if at all. This is because obtaining a model from quantities derived from the data, as opposed to directly from the data, is necessarily *equal to* or *worse than* obtaining the model from the data directly since derived quantities contain as much as or less information than the data themselves. For instance, fitting histogrammed data is an information inefficient and unreliable approach to obtaining models as it demands downsampling via binning and an arbitrary choice of bin sizes.

Besides information efficiency, obtaining models from unprocessed data also has another critical advantage that gets to the heart of scientific practice. While error bars around individual data points may be imperfectly known, they are, by construction, *better characterized* than error bars around derived quantities. Thus error bars around models determined from derived quantities are necessarily only as good as, but often less reliable, than error bars around models determined from the raw data. Unfortunately, as error bars around derived quantities can become too difficult to compute in practice, they are often ignored altogether. Nevertheless, error bars are a cornerstone of modern scientific research. They not only help quantify reproducibility but they also directly inform error bars around the models obtained and, as such, inspire the formulation of new competing models.

Putting it all together, it becomes clear that a model is *best informed*, and has the *most reliable error bars*, when learned from the data available in as raw a form as accessible from the experiments. This is true so long as it is computationally feasible to obtain models from such raw data and, as we will see in subsequent chapters, we are far from reaching computational bottlenecks in most problems of interest across the Natural Sciences.

1.1.2 Why do we formulate models with random variables?

If there is no uncertainty involved, a physical system is adequately described using deterministic variables. For example, Newtonian mechanics are expressed in terms of momenta, positions, and forces. However, when a system involves any degree of uncertainty, either due to noise, poor characterization of some or all of its constituents, features as of yet unresolved or otherwise fundamentally stochastic, then it is better described using *random variables*. This is true of the probabilistic nature of Quantum Mechanics as well as Statistical Physics and, as we illustrate herewith, also of Data Analysis.

Random variables are used to represent observations generated by stochastic systems. Stochasticity in Data Analysis arises due to inherent randomness in the physical phenomena of interest or due to measurement noise or both. Random variables are useful constructs because, as we will see, they are mathematical notions that reproduce naturally stochastic relationships between uncertain effects and observations; while, their deterministic counterparts cannot.

Note 1.2: Measurement noise

It is sometimes thought that models with probabilistic formulations are only required when the quantities of interest are inherently probabilistic. Nevertheless, measurement noise corrupts experimental observations irrespective of the quantities themselves being probabilistic or not. Consequently, probabilistic models are *always required* whenever models are informed by experimental output.

Random variables are abstract notions that most often represent numbers or collections of numbers. However, more generally, random variables can be generic notions that may include non-numeric quantities such as: labels for grouping data, *e.g.*, group A, group B; logical indicators, *e.g.*, true, false; functions, *e.g.*, trajectories or energy potentials. In all cases, numeric or not, random variables may be *discrete*, *e.g.*, dice rolls, coin flips, photon counts, bound energy states or *continuous*, such as temperatures, pressures or distances. Further, random variables may be finite collections of individual quantities, *e.g.*, measurements acquired during an experiment or even infinite ones, *e.g.*, successive positions on a *Brownian* particle's trajectory. At any rate, random variables have unique properties, which we will shortly explore, that allow us to use them in the construction and evaluation of meaningful probabilistic models.

Commonly, we imagine a random variable, which we denote with W, as being instantiated or assigned a specific value realized at w as a result of performing a measurement which amounts to a *stochastic event*. That is, we think of a measurement output w as a *stochastic realization* of W. Our stochastic events entail randomness inherited through W and influencing the assigned values w. We therefore distinguish between a random variable W and its realizations, w, *i.e.*, the particular values that W attains or may attain.

Stochastic events may encompass *physical* events, like the occurrence of chemical reactions or events in a cell's life cycle. Stochastic events may also encompass *conceptual* events, like an idealized version of a real-life system expressed in terms of fair coin tosses or, even, like instantaneously learning the spin orientation of a faraway particle given a local measurement of another spin to which the first is entangled.

Example 1.1: The photo-electric effect

When a photon falls onto certain materials, photo-electrons are sometimes emitted. Such a phenomenon provides the basis for a stochastic event.

In the photo-electric setting, it is often convenient to formulate a random variable W that counts the number of photo-electrons emitted. This random variable may take values $w = 0, 1, 2, \cdots$.

To develop a model, we imagine a *prototype experiment* as a sequence of stochastic events that produce N numeric measurements or, more generally, observations of any kind. We typically use w_n to denote the n^{th} observation and use $n = 1, \cdots, N$ to index them. As we highlighted earlier, individual observations in our experiment may be scalars, for example $w_n = 20.1$°C or $w_n = 0.74$ μm^3 for typical measurements of room temperature or an *E. coli*'s volume, respectively, or even non-numeric, such as $w_n =$ p.R83SfsX15 for descriptions of gene mutations. In general, we do not require that each observation in our experiment be of the same type; that is, w_1 may be a temperature while w_2 may be a volume.

As we will often do, we gather every observation conveniently together in a list

$$w_{1:N} = \{w_1, w_2, \cdots, w_N\}$$

and use subscripts $1:N$ to indicate that the list $w_{1:N}$ gathers every single w_n with an index n ranging between 1 and N. Unless explicitly needed to help draw attention to the subscript, for clarity, we may sometimes suppress this subscript and write simply w for the entire list.

As we have already mentioned, the observations $w_{1:N}$ are better understood as realizations of appropriate random variables $W_{1:N} = \{W_1, W_2, \cdots, W_N\}$ that we use to formulate our model.

1.1.3 Why do our models have parameters?

Models are mathematical formulations to which we associate parameters. Both models and their associated parameters are specialized to particular systems, experiment types and experimental setups. Assuming a model structure encoded in $W_{1:N}$ and provided observed values $w_{1:N}$, our main objective in Data Analysis becomes the estimation of the model's associated parameters.

Example 1.2: Normal random variables

The mean of a sequence of identical random variables W_n is only probabilistically related to each measured value w_n. For the simple example of a normally distributed sequence W_n, what we call the model is the Normal distribution, often also termed the Gaussian distribution. The associated parameters are: the mean μ and variance $v = \sigma^2$, with σ called the standard deviation, which indicate the center and spread of the values $w_{1:N}$, respectively. These are collectively described by the list of model parameters $\theta = \{\mu, v\}$. As illustrated in fig. 1.1, and as we will see in detail in later chapters, θ can be estimated from $w_{1:N}$.

In the previous example, the Gaussian forms a simple model that contains two parameters, namely the mean μ and the variance v, that we gather in θ. More generally, our models may contain K individual parameters that we may also gather in a list $\theta_{1:K} = \{\theta_1, \theta_2, \ldots, \theta_K\}$.

Typically, the parameters $\theta_{1:K}$ represent quantities we care to *estimate*, for example μ and v. A model is deemed *specified* when *numerical values* are assigned to $\theta_{1:K}$. Thus, specifying a model is understood as being equivalent to assigning values to $\theta_{1:K}$. Similarly, deriving error bars around the assigned values of $\theta_{1:K}$ is equivalent to deriving error bars around the model.

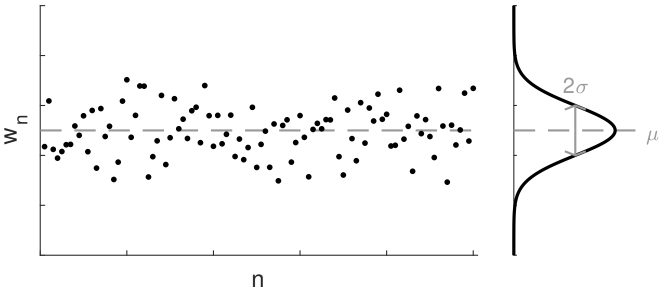

Figure 1.1: On the left hand side we show the output of an experiment after successive trials which we index with n. On the right hand side we find a histogram of the data with very fine bin sizes that assumes the shape of a Gaussian distribution. We denote the mean of this distribution by μ and standard deviation by σ.

As we invariably always face some degree of measurement noise, we formulate an experiment's results $w_{1:N}$ as probabilistically related to the parameters $\theta_{1:K}$. In the context of our *prototype experiment*, we incorporate such relations through the random variables $W_{1:N}$ and in the next section we lay down some necessary concepts.

Note 1.3: Modeling terminology

Strictly, by model in this chapter we mean the mathematical formulation itself alongside numerical values for its associated parameters. When we speak of measurements, observations, assessments, or data points we refer to the random variables $W_{1:N}$ and their realizations $w_{1:N}$. Similarly, by calibrating a model we imply selecting the correct values for its associated parameters (and sometime also characterizing their uncertainty). Determining both model parameters and their uncertainty is collectively referred to as *model estimation* or model training.

1.2 Working with random variables

Before we embark on specific modeling and estimation strategies, we begin by exploring some important notions that we need in order to work with random variables and the distributions from which they are sampled. That is, just as we can easily deduce derivatives and integrals of complicated functions by remembering a few simple rules of Calculus, we can similarly deduce probability distributions of complicated models by remembering a few simple rules of probability that we put forth in this section.

As we will soon start using random variables not only to represent measurements W, but also other relevant quantities of our model, we begin using R to label generic random variables.

1.2.1 How to assign probability distributions

In any model, a random variable R is *drawn* or *sampled from* some *probability distribution*. We label such a distribution with $\mathbb{P}$ and we write

$$R \sim \mathbb{P}.$$

In the language of Statistics this reads "the random variable R is sampled from the probability distribution $\mathbb{P}$" or "R follows the statistics of $\mathbb{P}$".

In statistical notation, in writing $R \sim \mathbb{P}$, we use $\mathbb{P}$ as a notational shorthand that summarizes the most important properties of the variable R. These include a description of the values r that R may take and a recipe to compute probabilities associated with these r's. As we will see in many cases, most often we work with probability distributions which are associated with probability density functions. In such cases, it is more convenient to think of $R \sim \mathbb{P}$ as a compact way of communicating: the allowed values r of R; and that these values obey the probability density $p(r)$ associated with $\mathbb{P}$.

Example 1.3: The Normal distribution

We previously encountered the Normal distribution, Normal(μ, v). A shorthand like

$$R \sim \text{Normal}\,(\mu, v)$$

captures the following pieces of information:

- The particular values r that R attains are real numbers ranging from $-\infty$ to $+\infty$.
- The probability density $p(r)$ of R depends on two parameters, μ and v, and has the form

$$p(r) = \frac{1}{\sqrt{2\pi v}} \exp\left(-\frac{1}{2}\frac{(r-\mu)^2}{v}\right).$$

Furthermore, the two parameters μ and v can be interpreted as the mean and the variance of R, respectively, since integration of the density leads to

$$\begin{aligned}(\text{Mean of } R) &= \int_{-\infty}^{+\infty} dr\, r p(r) = \mu, \\ (\text{Variance of } R) &= \int_{-\infty}^{+\infty} dr\, (r-\mu)^2 p(r) = v.\end{aligned}$$

Using the density $p(r)$, we can also compute the probability of measuring any value r between some specified $r_{\min}$ and $r_{\max}$. In particular, this is

$$\int_{r_{\min}}^{r_{\max}} dr\, p(r) = \frac{1}{2}\left[\operatorname{erf}\left(\frac{r_{\max}-\mu}{\sqrt{2v}}\right) - \operatorname{erf}\left(\frac{r_{\min}-\mu}{\sqrt{2v}}\right)\right] \tag{1.1}$$

where $\operatorname{erf}(\cdot)$ is the error function defined by an integral

$$\operatorname{erf}(r) = \frac{2}{\sqrt{\pi}} \int_0^r dr'\, e^{-\frac{1}{2}(r')^2}.$$

Example 1.4: The Exponential distribution

The Exponential distribution arises in many applications. A shorthand like

$$R \sim \text{Exponential}\,(\lambda)$$

captures the following pieces of information:

- The particular values r that R attains are real numbers ranging from 0 to ∞.
- The probability density $p(r)$ of R depends on one positive parameter, λ, and has the form

$$p(r) = \lambda e^{-\lambda r}.$$

The parameter λ can be interpreted as the reciprocal of the mean of R, since integration of the density leads to

$$(\text{Mean of } R) = \int_0^{\infty} dr\, r p(r) = \frac{1}{\lambda}.$$

Through the density $p(r)$, we can also compute the probability of measuring any value r between some specified $r_{\min}$ and $r_{\max}$. In particular, this is

$$\int_{r_{\min}}^{r_{\max}} dr\, p(r) = e^{-\lambda r_{\min}} - e^{-\lambda r_{\max}}. \tag{1.2}$$

Example 1.5: The multivariate Normal$_M$ distribution

The multivariate Normal$_M$ distribution is a generalization of the univariate Normal of example 1.3. In fact, the two definitions coincide for $M = 1$. The shorthand

$$\boldsymbol{R} \sim \text{Normal}_M\left(\boldsymbol{\mu}, \boldsymbol{V}\right)$$

captures the following pieces of information:

- The particular values $\boldsymbol{r}$ that $\boldsymbol{R}$ attains are real vectors of size M.
- The probability density $p(\boldsymbol{r})$ of $\boldsymbol{R}$ depends on two parameters, $\boldsymbol{\mu}$ and $\boldsymbol{V}$, and has the form

$$p(\boldsymbol{r}) = \frac{1}{\sqrt{(2\pi)^M |\boldsymbol{V}|}} \exp\left(-\frac{1}{2}(\boldsymbol{r}-\boldsymbol{\mu})\boldsymbol{V}^{-1}(\boldsymbol{r}-\boldsymbol{\mu})^T\right).$$

The parameter $\boldsymbol{\mu}$ is also a vector of size M and the parameter $\boldsymbol{V}$ is a positive definite square matrix of size M. Here, $|\cdot|$ is the matrix determinant. Similar to the univariate case, the two parameters $\boldsymbol{\mu}$ and $\boldsymbol{V}$ can be interpreted as the mean and the covariance of $\boldsymbol{R}$, respectively.

In the simplest case, a normally distributed bivariate random variable $\boldsymbol{R} = (R_1, R_2)$ may be written as

$$(R_1, R_2) \sim \text{Normal}_2\left((\mu_1, \mu_2), \begin{pmatrix} v_1 & \rho\sqrt{v_1 v_2} \\ \rho\sqrt{v_1 v_2} & v_2 \end{pmatrix}\right).$$

In this parametrization $\mu_1, \mu_2, v_1, v_2, \rho$ are scalars, v_1, v_2 are positive, and ρ is bounded between -1 and $+1$. In this case, the density takes the equivalent form

$$\begin{aligned} p(r_1, r_2) =& \frac{1}{2\pi\sqrt{v_1 v_2 (1-\rho^2)}} \\ & \times \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(r_1-\mu_1)^2}{v_1} + \frac{(r_2-\mu_2)^2}{v_2} - 2\frac{(r_1-\mu_1)(r_2-\mu_2)}{\sqrt{v_1 v_2}}\right)\right). \end{aligned}$$

Throughout this book, we extensively use several common distributions. In examples 1.3 and 1.4 we introduced two of them though many more are to come. As these will appear frequently, to refer back to them, we adopt a convention that we summarize in appendix B. Briefly, we use $R \sim \text{Normal}(\mu, v)$ and $\text{Normal}(\mu, v)$ to denote a Normal random variable and the Normal distribution, respectively. Furthermore, we use $\text{Normal}(r; \mu, v)$ to help distinguish this associated density with its distribution. According to our convention, the values r of the random variable R do *not* appear in the distribution $\text{Normal}(\mu, v)$; while, they *do* appear in the density $\text{Normal}(r; \mu, v)$. In the latter, we separate with ";" the variable values r from the parameters μ and v. We apply the same convention to the other distributions and densities as well.

As we distinguish between a random variable R and its values r, for clarity, in this chapter we also distinguish between a probability distribution $\mathbb{P}$ and its associated density $p(r)$. However, in subsequent chapters, we relax this convention whenever there is no ambiguity.

Distributions on random variables with probability density functions

For a random variable R whose distribution has a probability density, we can compute the probability of attaining any of the values gathered in η, where η is a collection of r values, by the integral

$$P_\eta = \int_\eta dr\, p(r). \tag{1.3}$$

In this integral, $p(r)$ is the *probability density function* of $R \sim \mathbb{P}$ and its precise form is characteristic of the distribution $\mathbb{P}$. For instance, as we have seen on examples 1.3 and 1.4, a Normal distribution $\text{Normal}(\mu, v)$ has a Normal density $p(r) = \exp\left(-\frac{(r-\mu)^2}{2v}\right)/\sqrt{2\pi v}$ and an exponential distribution $\text{Exponential}(\lambda)$ has an exponential density $p(r) = \lambda e^{-\lambda r}$. For these two, eq. (1.3) reduces to eqs. (1.1) and (1.2), respectively.

Note 1.4: Probability density over inadmissible values

By convention, the density $p(r)$ is assumed equal to zero over inadmissible values r. For example, if the random variable models a distance, $p(r) = 0$ for $r < 0$; similarly, if the random variable models a temperature reported in Kelvin units, $p(r) = 0$ for $r < 0$.

By definition, the area, or more generally the volume, underneath the entire probability density $p(r)$ must be equal to 1. This is called the *normalization condition* and implies that an η including *every* admissible value r has probability 1. For instance, from eqs. (1.1) and (1.2) we can see that the probabilities of sampling any real scalar value is equal to 1 for either Normal or Exponential random variables.

As can be seen from eq. (1.3), a density $p(r)$ is *unitful* and its units are determined by normalization. Since $\int dr\, p(r) = 1$, where the region of integration includes every admissible value, the density $p(r)$ has the *reciprocal units* of r. So, if r is a length (in cm), the density $p(r)$ has units of reciprocal length (1/cm); similarly, if r is a time (in s), the density $p(r)$ has units of frequency (Hz).

Note 1.5: Re-sampling the same value of a random variable

Equation (1.3) already signals that the probability of sampling a continuous scalar random variable between some values $r_{\min}$ and $r_{\max}$ is

$$P_{r_{\min},r_{\max}} = \int_{r_{\min}}^{r_{\max}} dr\, p(r). \tag{1.4}$$

This brings up an interesting point: there is a vanishingly small probability for the same value of a continuous scalar random variable to be sampled twice with finite samplings. In fact, the probability of sampling *any particular value* is 0 as can be seen by having coinciding $r_{\min}$ and $r_{\max}$ in the integral eq. (1.4). This indicates that, when thinking about probabilities over continuous variables, we need to consider *intervals* of values rather than isolated values.

This feature generalizes to any continuous random variable that need not necessarily be scalar; but, as we will see shortly, it does not carry over the values of discrete random variables which can re-occur even in finite samplings. For example, a roll of 4 will almost certainly re-occur multiple times in a total of 1000 rolls of a fair dice.

For a random variable R, it is also possible, and often useful, to *transform* its density $p(r)$ into a density $q(r')$ over another random variable R' with values related by a given function $r' = f(r)$. For example, such a transformation occurs when we want to apply a change to our coordinate system or simply otherwise re-parametrize our model. As we require the transformation $R \mapsto R'$ to leave unaffected the probabilities we compute either using the initial or transformed variables, the two densities must satisfy

$$\int_\eta dr\, p(r) = \int_{f(\eta)} dr'\, q(r').$$

In this equality, $f(\eta)$ is a collection that contains the transformed values $r' = f(r)$ of all r in the initial collection η. In the most general setting, it is hard to relate mathematically the densities $p(r)$ and $q(r')$ any further. However, provided $f(r)$ is a *differentiable* function that can be *inverted uniquely*, as is often the case, we may apply a change of variables on the right-hand-side integral to reach $\int_\eta dr\, p(r) = \int_\eta dr\, |J_{r\mapsto r'}| q(r')$. Here, $|J_{r\mapsto r'}|$ is the absolute value of the determinant of the transformation's Jacobian. In turn, since such an equality holds for any η, we may drop the integrals to reach a simpler form

$$q(r') = \frac{1}{|J_{r\mapsto r'}|} p(r). \tag{1.5}$$

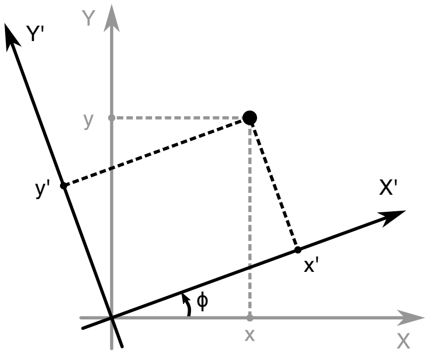

Figure 1.2: A random Cartesian position in the initial (X, Y) and the transformed (X', Y') frames of reference.

Example 1.6: Rescaling of random variables

Any physical quantity measured in real-life experiments almost always carries units. For practical reasons, often we need to convert between quantities reported in one system of units to another. Unit conversion itself is a simple example of variable transformation.

For concreteness, we consider a random variable R reported in some units and suppose $r' = \xi r$ where r' is expressed in different units from r. Here, ξ is the conversion factor, for example ξ could be 100 cm/m for r' expressed in terms of centimeters and r in terms of meters. In this example, both random variables R and R' are scalar, and so the Jacobian reduces to a simple derivative. More specifically, $|J_{r \mapsto r'}| = |df(r)/dr| = \xi$ and so, the densities are

$$q(r') = \frac{p(r)}{\xi}.$$

Example 1.7: Coordinate transformation of spatial random variables

Measurements of position are reported with respect to certain frames of reference. Changing the frame of reference is another example of a variable transformation.

For concreteness, we consider a bivariate random variable (X, Y) that models a location in the Cartesian plane, and suppose that (X', Y') is the same location in another Cartesian frame of reference rotated through an angle ϕ about the origin, see fig. 1.2. In this case, the original and transformed positions are related through

$$x' = x \cos\phi + y \sin\phi, \qquad y' = -x \sin\phi + y \cos\phi,$$

and the Jacobian of the transformation has the form

$$J_{(x,y)\mapsto(x',y')} = \begin{pmatrix} \frac{\partial x'}{\partial x} & \frac{\partial x'}{\partial y} \\ \frac{\partial y'}{\partial x} & \frac{\partial y'}{\partial y} \end{pmatrix} = \begin{pmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{pmatrix}.$$

Since $|J_{(x,y)\mapsto(x',y')}| = \cos^2\phi + \sin^2\phi = 1$, the densities in the two coordinate systems are

$$q(x', y') = p(x, y).$$

Distributions on random variables with discrete values

If $\rho_{1:M} = \{\rho_1, \rho_2, \dots, \rho_M\}$ gathers every admissible value of a *discrete* random variable R, then its probability density has the generic form

$$p(r) = \pi_{\rho_1}\delta_{\rho_1}(r) + \cdots + \pi_{\rho_M}\delta_{\rho_M}(r) = \sum_{m=1}^{M} \pi_{\rho_m}\delta_{\rho_m}(r) \tag{1.6}$$

where π_{ρ_m} are the probabilities of the individual values ρ_m contained in $\rho_{1:M}$. The *Dirac* terms $\delta_\rho(r)$ are specified by the properties

$$\begin{aligned} \delta_\rho(r) &= 0, \qquad\qquad\qquad\qquad r \neq \rho \\ \int dr\, \delta_\rho(r) &= 1 \end{aligned}$$

where the integral is taken over every allowed value of r; see appendix C.

Note 1.6: What is a discrete random variable?

One way to gain some intuition about discrete random variables is to consider *limiting* cases of continuous ones. For instance, we may consider acquiring measurements where we wish to distinguish between M distinct scalar values $\rho_{1:M}$. In a real-life experiment, our acquisitions are contaminated with noise and, for this reason, our measurements are generally scattered *around* the values $\rho_{1:M}$. As such, we may model our measurements with a random variable $R \sim \mathbb{P}$ which, due to the noise, attains continuous values; fig. 1.3.

In a noisy scenario, the scattering of r around $\rho_{1:M}$ is wide, and our measurements are found generally anywhere around and between $\rho_{1:M}$. In this case, a fine separation between the outcomes $\rho_{1:M}$ might be impossible. However, in increasingly clearer scenarios, our measurement distribution $\mathbb{P}$ concentrates around the outcomes giving rise to cleanly isolated peaks; fig. 1.3.

In the extreme limit of an idealized noiseless scenario, the distribution $\mathbb{P}$ places all of its probability around $\rho_{1:M}$ and its density, $p(r)$, becomes a train of Dirac terms as in eq. (1.6).

Normalization, in the case of a discrete random variable, reads $1 = \sum_{m=1}^{M} \pi_{\rho_m} \int dr\, \delta_{\rho_m}(r)$, where the integral over r spans any admissible and inadmissible value. Since probabilities π_{ρ_m} are dimensionless, this implies that each $\delta_{\rho_m}(r)$ on the right hand side of eq. (1.6) has dimensions of reciprocal r, similar to the density $p(r)$. As such, normalization of a discrete random variable's density can also take the equivalent form $1 = \sum_{m=1}^{M} \pi_{\rho_m}$.

One way to represent the distribution of a random variable with a density as in eq. (1.6) is

$$R \sim \text{Categorical}_{\rho_{1:M}}(\pi_{\rho_{1:M}})$$

where $\text{Categorical}_{\rho_{1:M}}(\pi_{\rho_{1:M}})$ denotes the Categorical *distribution* with outcomes $\rho_{1:M}$ and associated probabilities $\pi_{\rho_{1:M}}$. A random variable drawn from this distribution samples an outcome, ρ_m, in proportion to that outcome's probability, π_{ρ_m}; see fig. 1.4.

Example 1.8: Dice rolls modeled as Categorical random variables

Rolling a common dice leads to one out of six outcomes that we idealize as the faces marked with the numbers "1" through "6". Provided we identify these outcomes with the categories ρ_m, for $m = 1, \dots, 6$, we can model a dice roll as a Categorical random variable

$$R \sim \text{Categorical}_{\rho_{1:6}}(\pi_{\rho_{1:6}})$$

where the probability a face marked with "m", or category ρ_m, is π_{ρ_m}. As we know, fair dice have equiprobable faces, $\pi_{\rho_1} = \cdots = \pi_{\rho_6} = 1/6$; but loaded dice do not follow these probabilities.

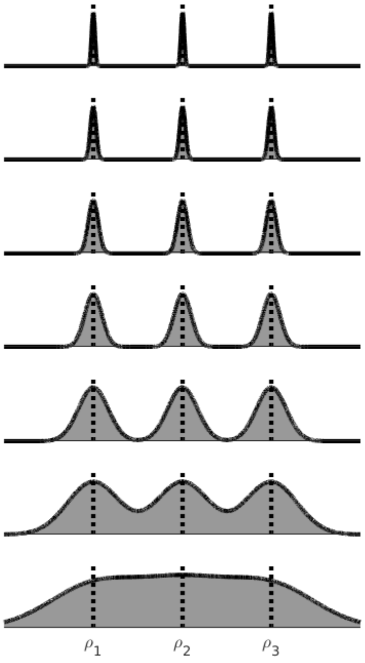

Figure 1.3: A discrete random variable arises as an idealization of increasingly less noisy measurements. Here, the bottom panel shows the probability distribution of highly noisy measurements. Upper panels show the probability distribution of successively clearer scenarios.

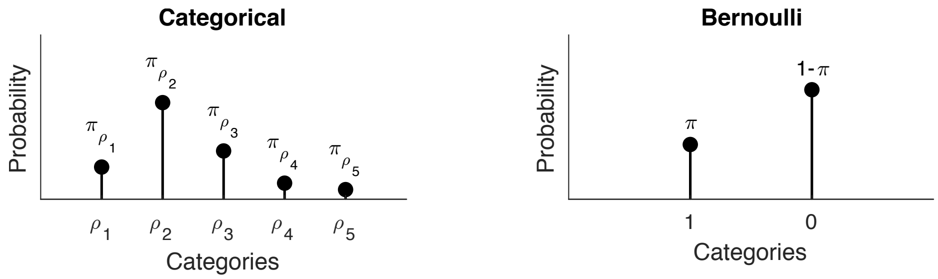

Figure 1.4: On the left hand side we plot the associated probabilities $\pi_{\rho_{1:5}}$ with 5 outcomes, $\rho_{1:5}$, of the Categorical distribution. To the right, we show a Bernoulli distribution, a special case of the Categorical distribution, with associated probabilities π and $1-\pi$ at its two possible outcomes.

The simplest example of a Categorical distribution is the Bernoulli distribution which is the special case having just two outcomes that conventionally we identify with the numbers 1 and 0, and respective probabilities $\pi_1 = \pi$ and $\pi_2 = 1-\pi$; see fig. 1.4. We often write Bernoulli(π) instead of the more elaborate Categorical$_{1,0}(\pi, 1-\pi)$.

Example 1.9: Coin flips modeled as Bernoulli random variables

An ideal coin flip has only two outcomes: “heads” or “tails”. Provided we identify these with the numbers 1 and 0, respectively, we can model a coin flip as a Bernoulli random variable

$$R \sim \text{Bernoulli}(\pi) \tag{1.7}$$

where π is the probability of “heads”. Here, specifying the probability of tails, $1-\pi$, is redundant since, by normalization, it is uniquely determined by π.

If, instead, we want to avoid identifying “heads” and “tails” with 1 and 0, we can also model a coin flip as a Categorical random variable

$$R \sim \text{Categorical}_{\text{heads,tails}}(\pi, 1-\pi). \tag{1.8}$$

Essentially, the only difference between eq. (1.7) and eq. (1.8) is in the meaning we assign to the values r, with the latter representation here having an *interpretational advantage* over the former.

Note 1.7: The Categorical and Bernoulli densities

An alternative to eq. (1.6) way of expressing a Categorical density is via the product

$$p(r) \propto \pi_{\rho_1}^{\Delta_{\rho_1}(r)} \cdots \pi_{\rho_M}^{\Delta_{\rho_M}(r)} = \prod_{m=1}^{M} \pi_{\rho_m}^{\Delta_{\rho_m}(r)}.$$

Here, the Kronecker terms $\Delta_\rho(r)$ are specified such that $\Delta_\rho(r) = 0$ when $r \neq \rho$ and $\Delta_\rho(r) = 1$ when $r = \rho$; see appendix C. In the special case of Bernoulli random variables, this reduces to

$$p(r) \propto \pi^r (1-\pi)^{1-r}.$$

As these expressions involve products as opposed to the sums of eq. (1.6), as we will see in subsequent chapters, these product forms are more convenient when we need to derive analytic formulas associated with discrete variables.

Distributions on random variables without probability density functions[*]

Since in subsequent chapters we will formulate models with random variables to which we *cannot* assign a probability density function, for example random variables that are smooth functions or trajectories or even random variables that are probability distributions themselves, we also need to account for appropriate distributions over these. In such cases, recipes to compute probabilities are case specific and, in general, the description of the associated distributions is necessarily more demanding. In examples 1.10 and 1.11 we provide only a sneak preview.

Example 1.10: The standard Brownian motion

We will examine Brownian motion in more detail in chapter 2. As we will see, standard Brownian motions in one dimension are random variables that represent functions from a time interval spanning 0 to some positive T to the real line. To denote them we write

$$X(\cdot) \sim \text{BMotion}_T^{\text{1D}}(D)$$

where the parameter D in the Brownian motion is a positive real scalar and, as we will see, can be interpreted as the diffusion coefficient of a particle diffusing in one dimension.

A shorthand like this captures the following pieces of information:

- The realizations of X are functions $x(\cdot)$ that, to any time t between 0 and T, assign $x(t)$ which is a position on the real line.
- Any realization of X, is initialized at the origin, *i.e.*, $x(0) = 0$.
- For any choice of times t and t' between 0 and T, the difference $x(t) - x(t')$ between the values $x(t)$ and $x(t')$ of any realization $x(\cdot)$ is a random variable itself.
- The random variable $x(t) - x(t')$ has a probability density given by

$$p\left(x(t) - x(t')\right) = \frac{1}{\sqrt{4\pi D|t-t'|}} \exp\left(-\frac{(x(t) - x(t'))^2}{4D|t-t'|}\right).$$

Example 1.11: The Gaussian process

We will examine *Gaussian processes* in more detail in chapter 6. As we will see, Gaussian processes are random variables that represent functions from a space S to the real numbers. To denote them we write

$$F(\cdot) \sim \text{GaussianP}_S\left(\mu(\cdot), C(\cdot,\cdot)\right).$$

A shorthand like this captures the following pieces of information:

- The realizations of F are functions $f(\cdot)$ that, to any point x in S, assign $f(x)$ which is a real number.
- The parameter $\mu(\cdot)$ is a function that, to every point x in S, assigns $\mu(x)$ which is also a real number.
- The parameter $C(\cdot,\cdot)$ is a function that, to every points x and x' in S, assigns $C(x, x')$ which is a non-negative real number.
- For any choice $x_1, \dots, x_M$ of any finite number M of points in S, the values $\boldsymbol{f} = (f(x_1), \dots, f(x_M))$ form a random array.
- The random array $\boldsymbol{f} = (f(x_1), \dots, f(x_M))$ has a multivariate Normal probability density given by

$$p(\boldsymbol{f}) = \text{Normal}_M\left(\boldsymbol{f}; \boldsymbol{\mu}, \boldsymbol{C}\right).$$

The parameters of this density depend upon the points $x_1, \dots, x_M$ and are given by

$$\boldsymbol{\mu} = \begin{pmatrix}\mu(x_1), & \cdots, & \mu(x_M)\end{pmatrix}, \qquad \boldsymbol{C} = \begin{pmatrix} C(x_1,x_1) & \cdots & C(x_1,x_M) \\ \vdots & \ddots & \vdots \\ C(x_M,x_1) & \cdots & C(x_M,x_M) \end{pmatrix}.$$

[*] This is an advanced topic and could be skipped on a first reading.

1.2.2 How to sample from probability distributions

So far we have discussed random variables and probability distributions from which random variables are drawn. What we discuss next is how to run *simulations*. That is, how to generate values or sample random variables in a computer from specified probability distributions. Random simulations are useful when we seek to re-create *in silico* repetitions of our prototype experiment. In subsequent chapters, we will see that we can use random sampling not only to re-create an experiment's results but also to draw inferences *from* an experiment's results.

Continuous random variables

For a random variable $R \sim \mathbb{P}$ that takes scalar real values r, its *probability cumulative function* also commonly termed the cumulative distribution function or, simply, CDF is a function $C(r)$ given by

$$C(r) = \int_{-\infty}^{r} dr' \, p(r') \tag{1.9}$$

where $p(r)$ is the probability density associated with $\mathbb{P}$. From this definition, we see that a CDF is dimensionless and increases monotonically between 0 to 1. This is a characteristic that we can use to develop a method from which to generate random values r of R on a computer.

For instance, starting with the density $p(r)$, we first calculate its CDF, $C(r)$. We then generate a random value, call it u, uniformly between 0 and 1, and ask: *for what value r is $C(r)$ equal to u?* In other words, we find $r = C^{-1}(u)$, where $C^{-1}(u)$ is the inverse function of $C(r)$. This method, often termed the *fundamental theorem of simulation*, is summarized in algorithm 1.1 and is visually illustrated in fig. 1.5.

Algorithm 1.1: Fundamental theorem of simulation for continuous variables

To simulate a continuous random variable $R \sim \mathbb{P}$:

- First, find the cumulative function $C(r)$ and its inverse $C^{-1}(u)$.
- Then, repeat the following steps:
  - Generate $U \sim$ Uniform$_{[0,1]}$.
  - Set $r = C^{-1}(u)$.

Upon completion, this algorithm generates values r according to $\mathbb{P}$.

Below we give a concrete example for an Exponential distribution.

Example 1.12: Simulating from an Exponential distribution

We consider a random variable $R \sim$ Exponential(λ). As we saw in example 1.4, this random variable takes real scalar values and its density is

$$p(r) = \begin{cases} \lambda e^{-\lambda r}, & r \geq 0 \\ 0, & r < 0 \end{cases}.$$

To apply the fundamental theorem of simulation, we first compute the CDF and its inverse. These are

$$C(r) = 1 - e^{-\lambda r}, \qquad C^{-1}(u) = -\frac{1}{\lambda}\log(1-u).$$

By means of algorithm 1.1, we then sample R as follows:

- First, sample

$$U \sim \text{Uniform}_{[0,1]}.$$

- Then, we compute r from

$$r = -\frac{1}{\lambda}\log(1-u). \tag{1.10}$$

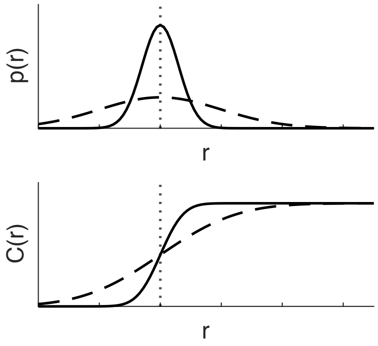

Figure 1.5: On the top panel we have PDFs broadly (dashed line) and tightly (solid line) centered around the same mean indicated by a vertical line. In the bottom panel, we have the corresponding CDFs. As seen, the CDF of the tighter PDF has a sharper slope near the mean, so in applying algorithm 1.1, most values of u would coincide with values of r near the mean. By contrast, the CDF of the wider PDF has a broader slope near its mean. In this case, the same values of u coincide with a larger range of values of r.

Since $\tilde{u} = 1 - u$ is also uniformly distributed between 0 and 1, for computational efficiency, when sampling Exponential random variables, we generate $\tilde{u} \sim \text{Uniform}_{[0,1]}$ in the first place and then use $r = -\frac{1}{\lambda}\log\tilde{u}$ instead of using eq. (1.10). In this way, we speed up the algorithm's execution by avoiding computation of the difference $1 - u$.

Note 1.8: Distribution functions

So far, we have encountered three important functions $p(r)$, $C(r)$, and $C^{-1}(u)$ associated with a distribution $\mathbb{P}$. These are very common in the literature and now we summarize some terms used to designate them:

- $p(r)$ is occasionally termed *probability density function* or *PDF*.
- $C(r)$ is occasionally termed *probability cumulative function, cumulative distribution function* or *CDF*.
- $C^{-1}(u)$ is occasionally termed *probability quantile function, inverse cumulative distribution function* or *ICDF*.

These functions may characterize the associated distribution $\mathbb{P}$. For this reason, they are often termed *probability distribution functions.*

Why the fundamental theorem of simulation for continuous variables works?*

In algorithm 1.1, we have a Uniform random variable u whose density is $g(u) = 1$ for any u between 0 and 1. When we set $r = C^{-1}(u)$, effectively we perform a transformation of random variables. As we saw earlier, the density $h(r)$ of the transformed variable is given by eq. (1.5) which, in this setting, takes the form

$$h(r) = \frac{1}{|J_{u\mapsto r}|}g(u).$$

Here, because both of our variables are scalar, the Jacobian of the transformation is given by the derivative

$$J_{u\mapsto r} = \frac{d}{du}C^{-1}(u) = \frac{1}{\frac{d}{dr}C(r)} = \frac{1}{p(r)}.$$

Considered together, the last two equalities lead to $h(r) = p(r)$. In other words, the values r generated in algorithm 1.1, indeed, follow the desired density $p(r)$.

*This is an advanced topic and could be skipped on a first reading.

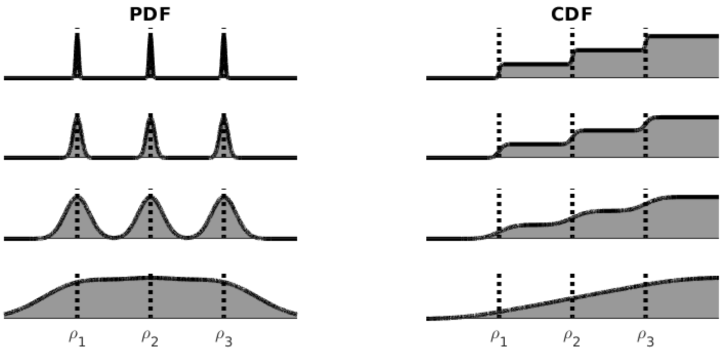

Figure 1.6: The CDF of a discrete random variable arises as an idealization of increasingly less noisy measurements. Here, the left panels show the probability distributions of noisy measurements and the right panels illustrate the corresponding CDFs.

Discrete random variables

We can use a similar procedure to sample discrete random variables too. In particular, for $R \sim \text{Categorical}_{\rho_{1:M}}(\pi_{\rho_{1:M}})$, the cumulative function is

$$C(\rho_m) = \sum_{m'=1}^{m} \pi_{\rho_{m'}}$$

or, more concretely, adopts the form

$$C(\rho_1) = \pi_{\rho_1}, \qquad C(\rho_2) = \pi_{\rho_1} + \pi_{\rho_2}, \qquad \cdots \qquad C(\rho_M) = \pi_{\rho_1} + \pi_{\rho_1} + \cdots + \pi_{\rho_M}.$$

To sample an outcome r, as with continuous random variables, we also need to generate $U \sim \text{Uniform}_{[0,1]}$. However, now a problem concerning the inversion of $C(r)$ arises. Namely, there may be no r such that $C(r) = u$. For this reason, instead of searching for outcomes such that $C(r) = u$, we search for the *lowest* value r such that $u \leq C(r)$. This version of the fundamental theorem of simulation is summarized in algorithm 1.2 and is visually illustrated in fig. 1.6.

Algorithm 1.2: Fundamental theorem of simulation for discrete variables

To simulate a random variable $R \sim \text{Categorical}_{\rho_{1:M}}(\pi_{\rho_{1:M}})$

- Generate $U \sim \text{Uniform}_{[0,1]}$.
- Find the lowest m such that $u \leq \pi_{\rho_1} + \pi_{\rho_2} + \cdots + \pi_{\rho_m}$.
- Set $r = \rho_m$.

At first, it might appear that this algorithm depends on the particular labeling of $\rho_{1:M}$ and that it would lead to different realizations r if the labels m had been assigned differently over the categories ρ_m. However, since relabeling of $\rho_{1:M}$ involves also a similar relabeling of $\pi_{\rho_{1:M}}$, this is *not* the case. In other words, this algorithm realizes each outcome $r = \rho_m$ with probability π_{ρ_m}, even when the labels m are reassigned over ρ_m.

Below we give a concrete example for a Bernoulli distribution.

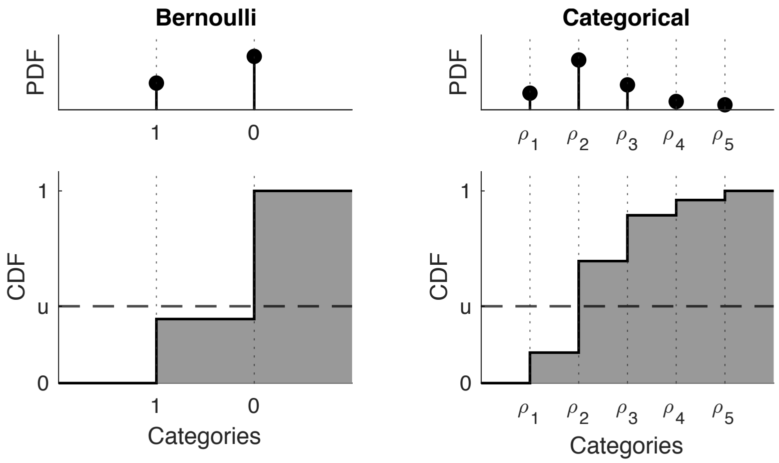

Figure 1.7: Here on the left panel, we consider a probability ranging from zero to one segmented into two portions of weight π and $1-\pi$ separated by a break point. We imagine these to be the probability of sampling outcome 1 or outcome 0 in a Bernoulli trial. To determine which outcome we select, we draw a Uniform random number u. In this figure, the u sampled falls above the break point. As such, we select outcome 0 for this Bernoulli trial. This argument carries over to the right panel with more than two portions.

Example 1.13: Simulation of Bernoulli random variables

Consider $R \sim$ Bernoulli(π). In this case, the cumulative function has a very simple form

$$C(1) = \pi, \qquad\qquad C(0) = 1.$$

To sample r, according to the fundamental theorem of simulation:

- First, we generate $U \sim$ Uniform$_{[0,1]}$.
- Then, if $u \leq \pi$ we set $r = 1$, else we set $r = 0$.

The two steps are illustrated in fig. 1.7.

Note 1.9: Intuition as to how the fundamental theorems of simulation works

There exists an intuitive explanation for the fundamental theorem of simulation that we illustrate in fig. 1.7. That is, we imagine an interval from 0 to 1 with a break point just as shown. The portion of the interval before the break point has length π and the remainder has length $1-\pi$. We now sample a Uniform random variable, u. If u falls before the break point, outcome 0 is realized. Otherwise outcome 1 is realized.

A similar logic holds for the Categorical distribution. Figure 1.7 shows the steps in the CDF of a discrete distribution. A draw from the Uniform distribution can be visualized as the dotted horizontal line. The value of the abscissa that coincides with the location where the dotted line intersects with the CDF dictates the value realized by the discrete random variable.

Why the fundamental theorem of simulation for discrete variables works?*

We consider the limiting scenario of note 1.6. Namely, we have a continuous random variable $R \sim \mathbb{P}$ in an experiment aiming to distinguish between the outcomes $\rho_{1:M}$. Generally, in a noisy experiment, $\mathbb{P}$ has a wide PDF and a CDF that increases smoothly from 0 to 1 as seen in fig. 1.6. In a less noisy experiment, however, the PDF has peaks and this means that the CDF retains smooth albeit clearly visible steps. As the noise level reduces,

*This is an advanced topic and could be skipped on a first reading.

the PDF's peaks become more prominent resulting in the CDF's steps becoming sharper. In the extreme case of a noiseless scenario, the CDF forms perfectly sharp steps mathematically represented by discontinuities. Seen as a limiting case of sampling continuous random variables, algorithm 1.2 is the direct analog of algorithm 1.1.

1.2.3 Manipulating probability distributions

In a prototype experiment, we most often have random variables with different properties. When handling such a complex model, we need to work simultaneously with more than one distribution. Here, we present bookkeeping rules that help us combine and manipulate multiple distributions.

Joint and marginal distributions

Provided the random variables in our model are independent from each other, for example because they may encode physical processes or observations that exert no influence upon each other, we may write

$$R_1 \sim \mathbb{P}_1, \qquad R_2 \sim \mathbb{P}_2, \qquad \cdots \qquad R_N \sim \mathbb{P}_N.$$

Each distribution $\mathbb{P}_1, \mathbb{P}_2, \ldots, \mathbb{P}_N$ is, in turn, associated with its own density $p_1(r_1), p_2(r_2), \ldots, p_N(r_N)$. As there is little chance of confusion, commonly we simply write $p(r_n)$ instead of $p_n(r_n)$.

Occasionally, we also encounter models with multiple random variables

$$R_1 \sim \mathbb{P}, \qquad R_2 \sim \mathbb{P}, \qquad \cdots \qquad R_N \sim \mathbb{P}, \tag{1.11}$$

that are independent and which also follow identical distributions $\mathbb{P}$, for example random variables that may model independent observations obtained from a time invariant system. On such occasions, we might abbreviate eq. (1.11) into $R_1, R_2, \ldots, R_N \overset{iid}{\sim} \mathbb{P}$ and speak of *independent and identically distributed*, or simply *iid*, random variables. Essentially, we mean that all densities $p(r_1), p(r_2), \ldots, p(r_N)$ happen to have the same form. In the iid setting, and only when there is no chance of confusion, we might refer to each one of the densities simply as $p(r)$. However, as we will see shortly, even with iid variables, it is often necessary to clarify what is meant by $p(r)$.

Following our convention, we denote the density of a single variable R_n as $p(r_n)$ and call it a *marginal density*. When multiple random variables $R_{1:N}$ arise in the same setting and a density gathers all of them, we write $p(r_{1:N}) = p(r_1, r_2, \ldots, r_N)$ and we refer to $p(r_{1:N})$ as a *joint density*.

A marginal density $p(r_n)$ is related to a joint density $p(r_{1:N})$ through an integration over the entire range spanned by $r_{1:n-1}$ and $r_{n+1:N}$. That is,

$$p(r_n) = \underbrace{\int dr_1 \cdots \int dr_{n-1} \int dr_{n+1} \cdots \int dr_N}_{\text{everything but } r_n} \, p(r_{1:N}). \tag{1.12}$$

Colloquially, we refer to the integration over variables, *i.e.*, from-right-to-left in eq. (1.12), as a *marginalization*. We refer to the reverse process, *i.e.*, from-left-to-right in eq. (1.12), as a *de-marginalization* or a *completion*.

Example 1.14: Marginalization

Similarly to how we obtain distributions over one random variable, we may obtain distributions over any subset of the variables in $R_{1:N}$. For concreteness, we consider a total of $N = 5$ variables and suppose that we wish to obtain a distribution over R_2 and R_4 only. In this case, marginal and joint densities are linked by

$$p(r_2, r_4) = \underbrace{\int dr_1 \int dr_3 \int dr_5}_{\text{everything but } r_2 \text{ and } r_4} \, p(r_{1:5}).$$

Now that we have discussed marginal and joint distributions, we are ready to discuss sampling of random variables from a Normal distribution.

Note 1.10: Box-Muller simulation of Normal random variables

We consider a random variable $X \sim \text{Normal}(\mu, v)$. As the CDF of X does not have a closed analytic form, we cannot use the fundamental theorem of simulation. For this reason, we follow a different approach. Namely, we start by considering two iid random variables

$$X, Y \overset{iid}{\sim} \text{Normal}(\mu, v).$$

The associated joint density reads

$$p(x,y) = p(x)p(y) = \left(\frac{1}{\sqrt{2\pi v}} \exp^{-\frac{(x-\mu)^2}{2v}}\right)\left(\frac{1}{\sqrt{2\pi v}} e^{-\frac{(y-\mu)^2}{2v}}\right).$$

On X and Y we perform three successive transformations:

- A linear transformation from x and y to

$$x' = \frac{x-\mu}{\sqrt{v}}, \qquad y' = \frac{y-\mu}{\sqrt{v}}.$$

- A non-linear transformation from Cartesian (x', y') to polar coordinates (ρ, ϕ) with

$$x' = \rho \cos\phi, \qquad y' = \rho \sin\phi.$$

- A non-linear transformation from ρ to λ with

$$\lambda = \rho^2.$$

The advantage of applying these transformations is that the resulting density over λ and ϕ is separable

$$p(\lambda, \phi) = \frac{1}{2}\exp\left(-\frac{\lambda}{2}\right)\frac{1}{2\pi} = \text{Exponential}\left(\lambda; \frac{1}{2}\right) \text{Uniform}_{[0,2\pi]}(\phi)$$

where we have made explicit that the Exponential and Uniform distributions are on the variables λ and ϕ, respectively.

The cumulative function over λ and ϕ can now be computed analytically. Thus, by generating two Uniform random samples $U_1, U_2 \overset{iid}{\sim} \text{Uniform}_{[0,1]}$, we can readily obtain random samples from the radial and polar angle distribution

$$\rho = \sqrt{\lambda} = \sqrt{-2\log u_1}, \qquad \phi = 2\pi u_2.$$

Transforming back to our original variables, we obtain

$$x = \mu + \sigma x' = \mu + \sigma\rho\cos\phi = \mu + \sigma\sqrt{-2\log u_1}\cos(2\pi u_2).$$

This algorithm for sampling Normal random variables is termed the *Box-Muller algorithm*. As can be seen, with little additional computational cost, this method also provides another Normal sample

$$y = \mu + \sigma\sqrt{-2\log u_1}\sin(2\pi u_2)$$

that is independent of x.

Conditional distributions

The order in which random variables arise in a model may be irrelevant, for example random variables modeling an experiment's observations that exert no influence upon each other, such as individual test scores, biometric measurements collected from a group of unrelated participants, or a temporal sequence of observations generated by a system at equilibrium. On the other hand, the order in which random variables arise *may be important*, for example random variables modeling observations of time-dependent phenomena, such as successive measurements of the number of cells in a growing cell culture, or the number of molecules available to react in a chain of chemical reactions.

To express *dependencies* among two random variables, R_1 and R_2, we write

$$R_2|r_1 \sim \mathbb{P}(r_1). \tag{1.13}$$

This reads "the random variable R_2, given the realization r_1 of the random variable R_1, is sampled from the probability distribution $\mathbb{P}(r_1)$" and means that the values of r_2 are associated with a density $p(r_2|r_1)$ depending upon r_1. We designate a distribution that depends upon the value of another random variable like $\mathbb{P}(r_1)$ and the associated density $p(r_2|r_1)$ as *conditionals*.

Note 1.11: How to avoid inaccuracies in specifying variable dependencies

In the setting of eq. (1.13), the random variable R_1 is sampled from its own (marginal) distribution that needs to be specified *separately*. In a complete model, both random variables $R_1 \sim \mathbb{P}_1$ and $R_2|r_1 \sim \mathbb{P}_2(r_1)$ need to be specified.

In a properly formulated model, the distribution of R_1 *must not* depend upon r_2 and, for this reason, the description of R_1 should precede that of $R_2|r_1$. If this is not possible, then we need to describe the random variables together through a single (joint) distribution $(R_1, R_2) \sim \mathbb{P}$.

Ideally, proper descriptions of models involving multiple random variables, that depend upon each other, should be provided in a nested fashion. For example

$$\begin{gathered} R_1 \sim \mathbb{P}_1 \\ R_2|r_1 \sim \mathbb{P}_2(r_1) \\ R_3|r_2, r_1 \sim \mathbb{P}_3(r_2, r_1) \\ \text{etc...} \end{gathered}$$

A necessary condition, although not always sufficient, for a reliable description of a probabilistic model, no matter how convincing the involved arguments may be, is that *every single distribution* $\mathbb{P}_1, \mathbb{P}_2(r_1), \mathbb{P}_3(r_2, r_1), \ldots$ be specified *clearly and explicitly*.

Whenever a model cannot be put in a nested form as above, even when random variables are grouped and joint distributions are applied, the model most likely contains flaws such as tautologies or contradictions. Consequently such a model is, even qualitatively, inappropriate.

In note 1.11, we consider nested variable dependencies. In particular, R_3 depends on the realizations r_2 and r_1, while R_2 depends on the realization r_1, and finally R_1 depends on no other realization. In the most general case, the probability distribution over our last random variable, say R_N, may depend on the realization of all previous random variables, $r_{1:N-1}$, and the same happens for all other variables up to the very first one r_1. On account of this hierarchy, resembling successive generations of variables, simulating a nested model requires a sampling algorithm termed *ancestral sampling* detailed below.

Algorithm 1.3: Ancestral sampling

```
To draw values for a group of random variables $R_{1:N}$, we proceed as follows:
  ▪ Find the density $p(r_1)$ associated with $R_1 \sim \mathbb{P}_1$.
  ▪ Sample $r_1$ using $p(r_1)$.
  ▪ For $n$ from 2 up to $N$, repeat:
      – Find the density $p(r_n|r_{1:n-1})$ associated with $R_n|r_{1:n-1} \sim \mathbb{P}_n(r_{1:n-1})$.
      – Sample $r_n$ using $p(r_n|r_{1:n-1})$.
```

Since we use ancestral sampling and hierarchical models extensively in the following chapter, we describe methods here to obtain the necessary conditional densities. Our starting point is the full joint density $p(r_{1:N})$ whose arguments we conveniently re-order and write as $p(r_{N:1})$.

Conditional and joint densities are related to each other through the *chain rule* which, in the most general setting, reads

$$p(r_{N:1}) = p(r_N|r_{N-1:1}) \cdots p(r_2|r_1)p(r_1).$$

In the simplest case, consisting of only two random variables, the chain rule reads

$$p(r_2, r_1) = p(r_2|r_1)p(r_1).$$

From this we immediately see that a conditional density over r_2 is normalized irrespective of r_1, *i.e.*,

$$\int dr_2 \, p(r_2|r_1) = \int dr_2 \, \frac{p(r_2, r_1)}{p(r_1)} = \frac{\int dr_2 \, p(r_2, r_1)}{p(r_1)} = \frac{p(r_1)}{p(r_1)} = 1.$$

Additionally, from the chain rule, we obtain two equalities, $p(r_2, r_1) = p(r_2|r_1)p(r_1)$ and $p(r_1, r_2) = p(r_1|r_2)p(r_2)$, that we can combine to obtain another important rule, namely *Bayes' rule*, which most often is written in the form

$$p(r_2|r_1) = \frac{p(r_1|r_2)p(r_2)}{p(r_1)}, \qquad\qquad p(r_1) \neq 0. \tag{1.14}$$

As we will see in subsequent chapters, eq. (1.14) is an indispensable tool in Data Analysis.

Example 1.15: Modeling dynamical systems

Dependency among variables is especially important when the physical system of interest evolves over time. In this dynamical setting, explored in depth in the next chapter, our prototype experiment is temporally structured: causality indicates that the last measurement W_N may be influenced by all preceding measured values, $w_{1:N-1}$; the penultimate measurement, W_{N-1}, may be influenced by all of its preceding ones $w_{1:N-2}$; and so forth.

With the rules of joint and conditional distributions, we can work out the densities of such models in the most general setting. For instance,

$$p(w_{1:N}) = p(w_N|w_{1:N-1})p(w_{N-1}|w_{1:N-2}) \cdots p(w_2|w_1)p(w_1).$$

It follows that if we need to sample realizations of $W_{1:N}$, we need 1 marginal and $N-1$ *different* conditional distributions. As a result, this sampling may become infeasible unless we make some assumptions.

- One drastic assumption, often too crude for realistic dynamical systems, is to assume that all variables are independent, which in this particular case is equivalent to assuming $p(w_n|w_{1:n-1}) = p(w_n)$. Under this assumption, the joint density factorizes into a product of densities

$$p(w_{1:N}) = p(w_1)p(w_2) \cdots p(w_N) = \prod_{n=1}^{N} p(w_n). \tag{1.15}$$

- Another, less drastic and often quite realistic, assumption is to consider $p(w_n|w_{1:n-1}) = p(w_n|w_{n-1})$. Under this assumption, the joint density also factorizes into a product of densities

$$p(w_{1:N}) = p(w_1)p(w_2|w_1) \cdots p(w_N|w_{N-1}) = p(w_1) \prod_{n=2}^{N} p(w_n|w_{n-1}). \tag{1.16}$$

Under these two assumptions, the total number of different probability distributions, that are needed to sample $W_{1:N}$, reduces from N in general to a single marginal, for eq. (1.15) or a marginal and a conditional, for eq. (1.16). This is under the assumption that the marginals, over each variable, and conditionals, over each pair of variables, are all the same.

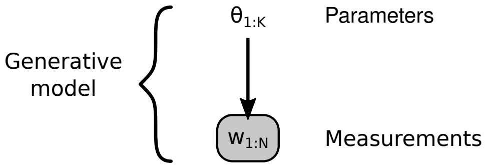

Figure 1.8: A generative model describes how measurements are generated. Implicitly, it encodes any influence the parameters $\theta_{1:K}$ exert upon the measurements $w_{1:N}$.

Somewhat pedantically, in deriving eq. (1.15), we invoked a so-called *0th order Markov assumption*

$$p(w_n|w_{1:n-1}) = p(w_n),$$

while, in deriving eq. (1.16) we invoked a so-called *1st order Markov assumption* often abbreviated simple as a *Markov assumption*

$$p(w_n|w_{1:n-1}) = p(w_n|w_{n-1}).$$

In principle, we can also invoke higher order assumptions where a measurement w_n is influenced by more than 1 past measurement; however, as we will see in the subsequent chapters, such assumptions are rarely used in practice, either because a 1st order assumption is already sufficient or because they lead to models with prohibitive computational cost.

1.3 Data-driven modeling and inference

Having introduced the necessary formalism, our emphasis from now on is not as much on mathematical rigor as it is focused on problem-formulation and problem-solving. But, *of what problem exactly?* With our basic notions laid down, we are now ready to define and address our problem more concretely.

In the data-centric context that is most appropriate for the Physical and Natural Sciences, we envision being provided information on a physical system such as:

- *How this system behaves* under relevant, well or poorly characterized, conditions.
- *How observations are acquired* on this system.
- *Specific values* of acquired observations.

These are the *data* and they serve as our input or starting point. Our primary task is to analyze the data and we tackle Data Analysis with the framework introduced in section 1.1.2. More specifically, within the framework set by the prototype experiment, which we adapt to real-life scenarios, our goal is to use the data to infer a model. However, before we can infer a model, we first need to go through a *synthesis stage* in order to develop the necessary mathematical formulation.

During the synthesis stage, we utilize the available information on our system to formulate the probability distribution $p(w_{1:N}|\theta_{1:K})$ that best describes our experiment. For example, in this stage we consider physical laws, dynamics, and noise properties, which, although non-numeric, in a very concrete sense are part of our given data. At this stage, we also decide on parameters $\theta_{1:K}$ and ascribe physical meaning to all or some of them.

The synthesis stage concludes with a concrete *generative model*; that is, a quantitive description of *how our experiment's measurements are generated*, see fig. 1.8. Our generative model mathematically links our unknowns, $\theta_{1:K}$, with our knowns $w_{1:N}$ and, in principle, could be simulated on a computer.

Note 1.12: A model's likelihood

The probability distribution $p(w_{1:N}|\theta_{1:K})$, mathematically established in a generative model, is a key quantity. This distribution is termed the *likelihood* or, colloquially, the likelihood function. The term follows from the notion that $p(w_{1:N}|\theta_{1:K})$ quantifies the likelihood of observing (sampling) the sequence of observations $w_{1:N}$ in our prototype experiment given the parameters $\theta_{1:K}$ which influence their realizations.

During the analysis stage, once we have formulated $p(w_{1:N}|\theta_{1:K})$, we apply the measured values of $w_{1:N}$ to compute parameter estimates, $\theta_{1:K}$. Traditionally, we call these values *estimators* and denote them with $\hat{\theta}_{1:K}$. We will see in chapter 3 that a likelihood provides us with a *universal* strategy to obtain $\hat{\theta}_{1:K}$ needed to specify uniquely a model we wish to learn. The challenge, however, is that we are also often interested in error bars around $\hat{\theta}_{1:K}$ or, put differently, probability distributions over all possible values of the random variable. For this reason, in chapter 4, we will consider an extended strategy that uses more than an experiment's likelihood.

The first stage in our workflow, namely setting up the generative model, constitutes a *modeling task*; while, the second stage, namely obtaining parameter estimates, constitutes a *computational task*. As we discuss in example 1.16, both stages in the solution of our problem are important and both stages pose unique challenges. As we will see in subsequent chapters, often we have to devise comprehensive approaches that deal with the challenges arising in both stages simultaneously.

Example 1.16: Likelihood based modeling and inference

As a concrete example, we imagine an experiment idealized as having one of two (discrete) measurement outcomes, for example the emission or not of a photo-electron as described in example 1.1. For simplicity, we may encode these outcomes with $\rho_1 = 1$ and $\rho_2 = 0$, respectively.

If we idealize individual assessments as iid, meaning that each measurement is independent of the others as in eq. (1.15), then the mathematical form of the likelihood is readily derived. In particular, the model responsible for generating the data takes the form

$$W_n|\pi \sim \text{Bernoulli}(\pi), \qquad n = 1 : N,$$

and, as of yet, has one unspecified parameter, namely π, to which we ascribe the meaning of probability that a single assessment results in a photo-electron emission. From now on, our goal is to estimate π.

To achieve our goal, we ask: *Given this generative model, what is the likelihood of our measurements?* This likelihood is the probability of observing the sequence $w_{1:N}$ and we may compute it as follows

$$p(w_{1:N}|\pi) = \prod_{n=1}^{N} p(w_n|\pi) = \prod_{n=1}^{N} \text{Bernoulli}(w_n;\pi) \propto \prod_{n=1}^{N} \pi^{w_n}(1-\pi)^{1-w_n} = \pi^M(1-\pi)^{N-M}$$

where we assumed that, within $w_{1:N}$, the first outcome, ρ_1, has been observed in total M times and the second outcome, ρ_2, has been observed the remainder of the times, namely $N - M$.

Finally, we estimate a value for the parameter π by asking: *Which value of π makes our observations most likely?* This is equivalent to asking which value of π makes $p(w_{1:N}|\pi)$ highest. Essentially, we seek the maximizer of $p(w_{1:N}|\pi)$. For instance, solving $\frac{d}{d\pi}p(w_{1:N}|\pi) = 0$, we recover $\hat{\pi} = M/N$, as intuitively expected.

For this example, we assumed that both outcomes, ρ_1 and ρ_2, are observed at least once and so $0 < \hat{\pi} < 1$, because $0 < M < N$. Yet had $M = 0$ or $M = N$, we may had erroneously concluded, due to limited data, that $\hat{\pi} = 0$ or $\hat{\pi} = 1$. Thus, even this toy example forebodes our need to go beyond approaches that rely exclusively on likelihoods.

In the Physical Sciences, data-driven approaches are sometimes termed *inverse methods*, *inverse problems*, or *inverse modeling*. Yet, as example 1.16 illustrates, there is nothing backward about obtaining models starting from the data and these, somewhat unfortunate terms, arose only because traditional approaches, namely obtaining models from the ground-up with a combination of first principles and data-fitting, came historically first and are now termed forward (or direct) methods.

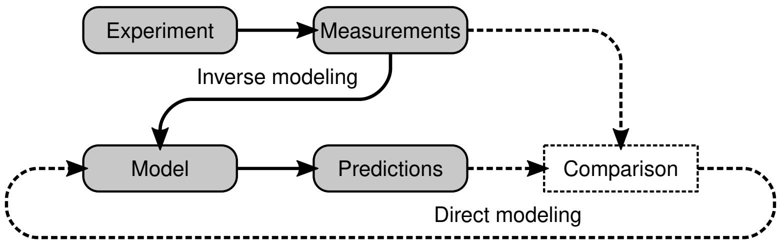

Figure 1.9: Here we show an illustration of the direct and inverse modeling paradigms. In the direct paradigm, a model is adjusted until its predictions agree with an experiment's measurements. By contrast, in the inverse paradigm, a model is inferred from the experiment's measurements without adjustments.

Note 1.13: Inverse modeling

Data-driven model inference essentially is an inverse problem. Solving an inverse problem is the opposite of solving a direct problem. Briefly, in a *direct problem*, also termed forward problem, we seek to determine an effect knowing its cause; while, in an *inverse problem* we seek to recover the cause knowing only the effect.

Inverse problems arise mainly when we need to interpret indirect physical measurements of unknown or partially known origin. For instance, when we are interested in elucidating the dynamics of complex biomolecules observed indirectly through fluorescence microscopy. In an experiment, we acquire images (measurements) with all sorts of artifacts that subsequently need to be removed in order to reveal the positions or dynamics of the biomolecules of interest. By contrast, simulating possible measurements (by, say, molecular dynamics simulations) and invoking a physical model, established or tentative, that *predicts* system behavior and subsequently checking whether predictions are in agreement or disagreement with the observed measurements constitutes direct modeling; see fig. 1.9.

A problem, whether direct or inverse, is *well-posed* when it meets the following conditions:

- The problem has a solution.
- The solution is unique.
- The solution does not differ substantially unless the supplied data also substantially differs.

These conditions are mathematically known as *existence, uniqueness,* and *stability*, respectively. If a problem fails to satisfy one or more of them, it is *ill-posed*.

Direct problems are well-posed when the effects (data) are well-defined, single-valued, and depend continuously on their causes. Often this is the case when we seek to reproduce observations mathematically or computationally. On the other hand, solutions to inverse problems do *not always exist*, or when they exist they are *almost never unique* or may *change dramatically* even when the supplied data (effects) differ only insignificantly. As a result, inverse problems are commonly ill-posed and solving them can be far more challenging.

Throughout the subsequent chapters, with the use of the appropriate random variables and probability distributions, we will see how inverse problems can be formulated statistically and how solutions to these problems can be computed robustly and efficiently.

Forward modeling, also termed direct modeling, has had its role to play and is heavily showcased throughout Physics where disparate observations were unified into predictive frameworks inspired by logic, symmetries and fundamental postulates. Undoubtedly, the forward approach has been tremendously successful. To wit, among others, it predicted the magnetic moment of the electron to a spectacular number of significant digits. But there are limitations to this historically successful approach.

While forward modeling historically came first, inverse methods, spurred in equal parts by advances in probability theory and motivating data-centric questions in the Natural Sciences, transiently gained prominence in mainstream Physics thanks to Laplace. Today, large swathes of complicated enough physical and chemical systems, in addition to Life and Social Sciences, are not naturally modeled from the ground-up, *i.e.*, starting from first principles. Instead, in these cases, observations often only suggest loose couplings between variables of interest and probabilistic relations between various quantities implied by the data.

The forward approach is different from the philosophy we adopt in this textbook. Instead, we use first principles only to motivate forms for our generative models. Beyond this, we are motivated by the practice of Statistics that instead attempts, from the onset, to be as agnostic about the model parameters (or the model itself) as possible and learn parameters and models self-consistently from the available data as efficiently as computationally possible.

1.4 Exercise problems

Exercise 1.1: Product of Normal densities

Show that the product of Normal probability densities remains a Normal probability density.

Exercise 1.2: Calculus warm-up

Evaluate the following.

1. Gaussian integral: $\int_{-\infty}^{+\infty} dx\, e^{-(x-\mu)^2/(2v)}$, assume μ and v are real scalars and $v > 0$.
2. Gaussian moments: $\int_{-\infty}^{+\infty} dx\, x^2 e^{-(x-\mu)^2/(2v)}$, assume μ and v are real scalars and $v > 0$.
3. Gaussian convolution: $\int_{-\infty}^{+\infty} dx e^{-(y-x)^2/(2v)} e^{-(x-\mu)^2/(2v)}$, assume μ, y and v are real scalars and $v > 0$.
4. Gamma function integral: $\int_0^{\infty} dx\, x^n e^{-x/a}$, assume n is a positive integer and $a > 0$.
5. Poisson variance: $\sum_{n=0}^{\infty} n^2 \lambda^n \exp(-\lambda)/n!$, assume $\lambda > 0$.

Exercise 1.3: Matrix algebra warm-up

Complete the following matrix algebra operations and, in doing so, state the conditions on the individual matrices (A, B, C and D) required for the operation to be well-defined. Further assume that all matrices and vectors are of the appropriate dimension to support the operations required.

1. Derive the matrix transpose relation: $(AB)^T = B^T A^T$ where T denotes the transpose.
2. Demonstrate the following through Taylor expansion: $e^A e^A = e^{2A}$.
3. Complete the squares in f: $fAf^T - fB^T - Bf^T$ where f is a row vector.
4. Perform the following matrix Normal convolution: $\int e^{(f-\mu)A^{-1}(f-\mu)^T} e^{(h-f)B^{-1}(h-f\mu)^T} df$ were f, μ and h are vectors with real coordinates and the integral is over every coordinate of f taken over the whole real line.
5. Verify the "inverse of sum" identity: $(A+B)^{-1} = A^{-1} - A^{-1}(B^{-1} + A^{-1})^{-1}A^{-1}$.
6. Verify the "push through" identity: $(\mathbb{1} + AB)^{-1} = \mathbb{1} - A(\mathbb{1} + BA)^{-1}B$.
7. Verify the Woodbury identity: $(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}$.

Exercise 1.4: Permutations and combinations

Consider integers $N = 1, 2, \ldots$ and $M = 0, 1, \ldots, N$.

1. Show that the total number of distinct arrangements (permutations) of M objects selected out of N distinct objects is $\frac{N!}{(N-M)!}$.
2. Show that if we ignore the arrangement of the objects (combinations) the total number drops to $\frac{N!}{M!(N-M)!}$.
3. Show that the total number of different combinations of N distinct objects is 2^N.

Exercise 1.5: Cumulative probability function

Explain why the cumulative probability function in eq. (1.9) takes only values between 0 and 1.

Exercise 1.6: Cumulative and quantile functions of exponential random variables

Verify the formulas of $C(r)$ and $C^{-1}(r)$ in example 1.12.

Exercise 1.7: Joint distribution

Show that a joint distribution encodes all information required to re-construct all relevant marginals and conditionals. For concreteness, consider a model with three random variables $R_{1:3}$.

Exercise 1.8: Sum of random variables

Consider two independent random variables R_1 and R_2 with densities $p_1(r_1)$ and $p_2(r_2)$, respectively. Use a transformation to show that the density $p_3(r_3)$ of a random variable R_3, with values $r_3 = r_1 + r_2$, is equal to the convolution $p_3(r_3) = (p_1 * p_2)(r_3)$.

Exercise 1.9: Minimum of Exponential random variables

Computing the minimum of two Exponential random variables is often relevant when considering the time of arrival of the first event given two competing events. For this reason here, we consider two exponential random variables $R_1 \sim$ Exponential(λ_1) and $R_2 \sim$ Exponential(λ_2). Show that the random variable R_3, with values $r_3 = \min(r_1, r_2)$, follows an Exponential$(\lambda_1 + \lambda_2)$ distribution.

Exercise 1.10: A sanity check on random variable rescaling

Verify that the density $q(r')$ of the rescaled random variable R' in example 1.6 has the correct units and that it is properly normalized.

Exercise 1.11: Linear transformations

In examples 1.6 and 1.7 we have seen how to obtain the probability density of random variables under rescaling and rotation. These two separate operations can be combined into a single one. For instance, consider a bivariate random variable (X, Y) and suppose that (X', Y') is the random variable under a linear transformation

$$x' = Ax + By + C, \qquad y' = Dx + Ey + F$$

where A, B, C, D, E, and F are constants. Find the probability density of (X', Y') in terms of $p(x, y)$ and to avoid degeneracies consider only the case with $AE - BD \neq 0$.

Exercise 1.12: Division of random variables

Consider a random variable R with values $r = 1/x$, where X is a scalar random variable with density $p(x)$. Compute the density $q(r)$ in terms of $p(x)$.

Exercise 1.13: Spherical coordinate transformations

Consider a tri-variate random variable (X, Y, Z) that models a position in Cartesian space. Use a transformation of random variables to relate the probability density $p(x, y, z)$ with the probability density $q(r, \phi, \theta)$ of the same position in spherical coordinates (R, Φ, Θ).

Exercise 1.14: Gamma random variables and derivatives

Suppose $R_1 \sim$ Gamma(α_1, β) and $R_2 \sim$ Gamma(α_2, β) are independent Gamma random variables; see appendix B for the definition of the Gamma distribution. Find the probability densities of the random variables V_1, V_2, V_3 with values

$$v_1 = r_1 + r_2, \qquad v_2 = \frac{r_1}{r_1 + r_2}, \qquad v_3 = \frac{r_1}{r_2}.$$

Exercise 1.15: Bernoulli random variables

Show that the sum of Bernoulli random variables is distributed according to a Binomial distribution; see appendix B for the description of a Binomial distribution if needed.

Exercise 1.16: The Gibbs inequality

Consider two Categorical probability distributions over the same categories $\rho_{1:M}$, one with parameters $\pi_{\rho_{1:M}}$ and the other with parameters $\pi'_{\rho_{1:M}}$. Prove the Gibbs inequality

$$-\sum_{m=1}^{M} \pi_{\rho_m} \log \frac{\pi'_{\rho_m}}{\pi_{\rho_m}} \geq 1.$$

Hint: use the fact that $\log x \leq x - 1$ for $x > 0$.

Exercise 1.17: Poisson convolution

Show that the sum of Poisson random variables remains a Poisson random variable. Hint: If necessary, see appendix B for the description of a Poisson distribution

Exercise 1.18: Manipulating transformed densities

Consider iid random variables R_1, R_2, R_3 with a common density $p(r)$. Assume R_1, R_2, R_3 are positive real scalar random variables. Find, in terms of $p(r)$, the probability that the polynomial $r_1 x^2 + r_2 x + r_3$ has real roots.

Exercise 1.19: A fair dice

Use the fundamental theorem of simulation to simulate a roll of a fair dice. Generate several rolls and verify that indeed the dice simulated is fair.

Exercise 1.20: The Weibull distribution

Consider a Weibull random variable X. This variable takes real scalar values and its probability density reads

$$p(x) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^{\alpha}}$$

for appropriate values α and β.

1. Describe and implement an algorithm that uses the fundamental theorem of simulation to simulate the random variable X.
2. Use your algorithm to generate a large number of random realizations x of X.
3. Histogram your output and compare to the analytic form of the Weibull density.

Exercise 1.21: Label invariance of the fundamental theorem of simulation

1. Apply the fundamental theorem of simulation to simulate draws from Categorical$_{\rho_1,\rho_2,\rho_3}(\pi_{\rho_1},\pi_{\rho_2},\pi_{\rho_3})$.
2. Verify that ρ_1,ρ_2,ρ_3 are realized with probabilities $\pi_{\rho_1},\pi_{\rho_2},\pi_{\rho_3}$, respectively.
3. Apply a relabeling of ρ_1,ρ_2,ρ_3 and verify that the fundamental theorem of simulation keeps yielding realizations with the correct probabilities.

Exercise 1.22: Normal random variables

1. Implement the Box-Muller algorithm of note 1.10 and generate a large number of Normal(μ,v) random values.
2. Use your generated values to construct histograms and verify that your implementation yields the correct statistics.

Exercise 1.23: Poisson random variables

1. Show that if the time between successive events is exponentially distributed, with rate λ, then the number of events expected within any time interval, of duration T, is distributed according to Poisson(μ) where $\mu = T\lambda$.
2. Develop a method to draw samples from a Poisson distribution by leveraging the fact that the time between events is exponentially distributed. Repeat the exercise for various values of λ, histogram your samples and compare your histogram to the coinciding Poisson density.

Exercise 1.24: A loaded dice

A dice is rolled 120 times yielding the results:

face	"1"	"2"	"3"	"4"	"5"	"6"
number of appearances	15	34	18	19	19	15.

Reason, based on likelihoods, that the dice is loaded.

Exercise 1.25: Random variable convolutions: the instrument response function

Consider a single photon detector recording exponentially distributed photon inter-arrival times, with rate λ. Detector electronics add a stochastic delay to the photon detection time distributed according to a Normal distribution with mean μ and variance v. Find the resulting probability distribution over photon detection times.

Project 1.1: The point spread function in fluorescence microscopy

In fluorescence microscopy, photons are detected at positions that differ probabilistically from the point at which they are emitted. In particular, under ideal imaging conditions, each photon emitted from a position $(x_\star, y_\star)$ is detected independently of the other photons at a position (X,Y) that is randomly distributed according to the *Airy probability density*

$$p(x,y) = \frac{4\pi n_\alpha^2}{\lambda^2}\left(\frac{J_1\left(\frac{2\pi n_\alpha}{\lambda}\sqrt{(x-x_\star)^2+(y-y_\star)^2}\right)}{\frac{2\pi n_\alpha}{\lambda}\sqrt{(x-x_\star)^2+(y-y_\star)^2}}\right)^2$$

where λ is the photon's wavelength, n_α is the microscope's numerical aperture and $J_1(\cdot)$ is the 1st Bessel function of the first kind. Typical values for the parameters are $\lambda = 510$ nm and $n_\alpha = 1.40$.

1. Verify that the probability density $p(x,y)$ is properly normalized.
2. Apply a transformation from Cartesian to polar coordinates and change the photon detection position from (X,Y) to radius and azimuth (R,Φ) relative to $(x_\star,y_\star)$.
3. Use the Airy density of (X,Y) to derive the density of (R,Φ).
4. Verify that the Airy density is radially symmetric.

5. Evaluate the Airy density at a set of fixed grid point at radii $r_{1:M}$.
6. Use your tabulation of the Airy density at these radii and numerical integration to approximate the CDF $C(r_m)$ at the grid's radii $r_{1:M}$.
7. Use interpolation to approximate the CDF $C(r)$ at radii between $r_{1:M}$.
8. Use the fundamental theorem of simulation and your interpolated $C(r)$ to simulate the detection position of a large number of photons.
9. Summarize your simulated positions in a histogram and verify that indeed your implementation produces photon detections from the correct distribution.

Project 1.2: EMCCD signal amplification

Light detectors based on *electron multiplication charge coupled devices* (EMCCD) are widely used in both telescope and microscope cameras. EMCCDs perform well under low light conditions as they amplify the detected signal. Signal amplification by an EMCCD is modeled as follows:

- A light source emits a Poisson distributed number of photons n_ϕ which strike the detector at a rate λ for a period τ_{exp}.
- Each of these photons may induce the transport of an electron into the electron multiplication (EM) register with probability q independently of the other photons resulting in a total of n_e electrons transported into the EM register.
- The electrons in the EM register are subsequently multiplied through an electron cascade process, which is well approximated by a Gamma distribution with shape n_e and scale G, outputting n_o electrons that are transferred to the analog-to-digital (A/D) converter.
- Due to thermal noise, the final readout w, resulting from A/D conversion, is normally distributed around n_o.

The full generative model is provided below:

$$\begin{aligned}
N_\phi|\lambda \sim & \text{Poisson}\left(\lambda\tau_{\text{exp}}\right),\\
N_e|n_\phi, q \sim & \text{Binomial}\left(q n_\phi\right),\\
N_o|n_e, G \sim & \text{Gamma}\left(n_e, G\right),\\
W|n_o, v \sim & \text{Normal}\left(n_o, v\right).
\end{aligned}$$

1. Derive the joint distribution over N_e and N_ϕ.
2. Derive $p(w, n_\phi, n_e, n_o|\lambda, q, G, v)$ and, from this density, compute $p(w|\lambda, q, G, v)$ by marginalization.
3. Explain, in words, your reasoning justifying which of the two likelihoods above you would hypothetically maximize in order to determine λ.

Additional Reading

C Bishop. Pattern recognition and machine learning. Springer, 2006.
DS Sivia, J Skilling. Data analysis: a Bayesian tutorial. Oxford University Press, 2006.
JA Rice. Mathematical statistics and data analysis. 3rd edition. Duxbury, 2007.
L Wasserman. All of Statistics: a concise course in statistical inference. 2nd printing. Springer, 2005.
M Hirsch, RJ Wareham, ML Martin-Fernandez, MP Hobson, DJ Rolfe. A stochastic model for electron multiplication charge-coupled devices–from theory to practice. PloS one, 8:e53671, 2013.