PART I

CONCEPTS FROM MODELING, INFERENCE, AND COMPUTING

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 into 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 preprocessing, 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 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, or 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 whether the quantities themselves are 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 nonnumeric 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 infinite quantities, *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 that 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 photoelectric effect

When a photon falls onto certain materials photoelectrons are sometimes emitted. Such a phenomenon provides the basis for a stochastic event.

In the photoelectric setting, it is often convenient to formulate a random variable W that counts the number of photoelectrons emitted. This random variable may take values $w = 0, 1, 2, \ldots$.

To develop a model, we imagine a *prototype experiment* as a sequence of stochastic events that produce N distinct numeric measurements or, more generally, observations of any kind. We typically use w_n to denote the nth observation and use $n = 1, \ldots, N$ to index them. As we highlighted earlier, individual observations in our experiment may be scalars, for example $w_n = 20.1°\text{C}$ or $w_n = 0.74\ \mu\text{m}^3$ for typical measurements of room temperature or an *E. coli*'s volume, respectively, or even nonnumeric, 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,

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

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 $W_{1:N} = \{W_1, W_2, \ldots, W_N\}$ 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 termed the Gaussian distribution. The associated parameters are the mean μ and variance $v = \sigma^2$, with standard deviation σ, that indicate the center and spread, respectively, of the values $w_{1:N}$. 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 Example 1.2, 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.

As we invariably 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.

Fig. 1.1 (Left) We show the output of an experiment after successive trials that we index with n. (Right) 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 the standard deviation by σ.

Note 1.3 Modeling terminology

In this chapter, when we use the term model, 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 them. As we will see, most often we work with probability distributions that 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 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

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

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

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

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

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

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

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

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

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

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