Physics 344
Simulation and Inference in Stochastic Systems
Theory Notes 2019

HC Eggers

Contents

Preface

- Starting in 2019, this course is undergoing extensive revision based on the changed needs and perceptions of computation and physics. As will become clear in these course notes, deduction and induction are central to scientific thinking and practice. In computation, these are most easily recognised in the tasks of *simulation*, which is deductive, and *inference*, which is inductive.
- The resulting extensive restructuring includes the following aspects:
  - *Simulation* has always been and remains a core component. Because it is best practised directly by computer coding, these lecture notes provide only the basic theoretical concepts and tools of simulation, leaving their implementation to the computational tasks and projects in the computer labs and workbooks. The basic tools of simulation include, of course, the standard algorithms and approaches, which in the present stochastic context imply the use of *Monte Carlo methods*. An important and often-overlooked second toolset is that of *dimensional analysis* based on the identification of *physical scales*.
  - With regard to *inference*, the new lecture notes focus almost exclusively on the Bayesian approach, leaving the traditional but widespread frequentist practices to the reader to experience elsewhere. The Bayesian framework receives, however, only cursory treatment because of time constraints and the present focus on a practical introduction. The development of the full power of Bayesian thinking is treated more fully in the Honours course *Bayesian Physics*.
  - Common to both simulation and inference are the basic tools of *statistics*. Along with pure mathematics and the mathematical part of pure computer science, statistics is neither deductive or inductive but instead a part of the language of logic. Much of the basic statistics retains its place in this course.
  - Much greater emphasis is placed on stochastic processes which had previously been taught in Physics 719. Correspondingly, the introduction to Monte Carlo methods is being expanded and brought into the context of both simulation and inference.
  - Due to time constraints and the additional content introduced, most of the theory of generating functions and the Central Limit Theorem has to be omitted.
- Throughout, the focus falls on a practical but mathematically grounded introduction to the basic concepts and tools. Compared to standard data science courses, there is much more emphasis on crucial concepts and mathematical skills rather than on methods and tricks. On the other hand, the maths provided falls well short of the fully topological and measure-theoretic approaches found in the pure statistics literature. We prefer pictures and analogies to formal definitions and theorems as a means to convey ideas.
- As the course notes are evolving strongly, they will be updated and made available in successive versions in PDF format. It is therefore probably better to either not print at all or to wait until a particular section has been treated fully in the lectures and tutorials.
- Corrections of errors and questions regarding these notes are always welcome.

Chapter 1

Foundations

1.1 Deduction and induction

- The new title of this course is *Simulation and Inference in Stochastic Systems*. It reflects the significant evolution in understanding since the inception of this course as well as the insight that quantitative science rests on the fundamental duality and mutual interplay of *deduction* and *induction*. While there is much to say about deduction and induction due to their foundational role, the goals of this course are more practical. For the purposes of this course, it is sufficient to provide a basic understanding.

- Knowledge consists of a set of *hierarchies*. At the top of each hierarchy is some generality; it can be a mathematical law, a rule, algorithm, concept in language. That law or concept then results in many different possible facts or observables which are at the bottom of the hierarchy.

- The facts or observations *follow* from the one law or rule. The process of *following* could be either physical causation or explanation of consequences or implications. Here are some examples:

  - A *flying intercontinental nuclear missile* as a concept with properties such as typical speed, trajectory, thermal emission intensity and spectrum, radioactivity etc is at the top of a hierarchy; the many cases of specific missiles, trajectories and the measurements of such are at the bottom.
  - A *house break-in* would have typical properties such as broken glass, door or window, an alarm system going off, missing items, suspicious people and vehicles etc.
  - An short-lived *elementary particle* like a pion or other meson can decay into two long-lived charged particles whose momentum and energy can be measured in a detector. Since the elementary particle has a fixed rest mass, plotting the invariant masses of resulting charged particle pairs will show a bump at the correct elementary particle mass. The particle and its physics represent the top, the specific measurements the bottom of this hierarchy.
  - Another typical physics example is Newton's Second Law, $\boldsymbol{F} = m\, d^2\boldsymbol{x}/dt^2$. This one law is applied to a multitude of different physical systems by using different force laws $\boldsymbol{F}(\boldsymbol{x})$ such as the gravitational force or electromagnetic force. Implemented with different initial conditions and parameter values, Newton II generates many predictions, data points, or facts which can be verified or observed in the real world.

  In all cases we have *One law; many possible resulting facts.*

- *Deductive reasoning and practice* starts out with the hypothesis of one particular "truth" and works out possible consequences in terms of actual or possible facts.

- *Inductive reasoning and practice* starts with an actual set of observations or facts or data. On the present level of discourse, the reality and existence of these facts is clear and not in doubt. Induction attempts to reconstruct the general law from which these facts followed. Unlike deduction, however, induction involves conjecture: how could the observer be *certain* that the general law which he inferred from the facts was indeed the correct one? What basis could he have for a claim that there is not an alternative one?

- *One hierarchy or many?*

  If the law or rule would indeed be unique, then there would be nothing further to do; one would merely work out the deductive consequences. Very often, however, the law is not unique: two or more possible laws may be able to explain the same set of facts, meaning that there is not one hierarchy but two or possibly more. Indeed, the abovementioned dilemma of inductive reasoning shows that there is room for doubt whether any law can be unique, to the total exclusion of alternatives.

  The easiest resolution to resolve competing explanations is to find a fact which is *not* possible or explainable within one of the laws, which thereby eliminates that law. *Falsification* is a valuable tool of science. Mostly, however, the available data does not result in such clear-cut conclusions. Rather than eliminating one competing hypothesis altogether, the available facts only helps us to decide how *likely* each one is.

- Faced with two or more alternative hypotheses, a purely deductive approach could try to successively use each as the starting point to arrive at the results. In that case, if Hypothesis $\mathcal{H}_A$ results makes facts $\mathcal{D}$ more likely than Hypothesis $\mathcal{H}_B$, a purely deductive approach could claim that thereby $\mathcal{H}_A$ is more likely than $\mathcal{H}_B$. This approach, called *maximum likelihood*, is logically flawed. The arrow of reasoning in deduction leads from $\mathcal{H}$ to $\mathcal{D}$ and not from $\mathcal{D}$ to $\mathcal{H}$, and inverting the direction of reasoning is logically inconsistent.

  For example: If it rains, the ground is very likely to be wet. Wet ground does not, however, necessarily imply that it is very likely raining.

- The failure of purely deductive thinking to provide a logically consistent procedure forces us to take seriously that laws, rules and explanations are not simply true or false; they can be more or less likely, given the data at hand. This may seem contradictory too, because only one of those laws would have actually resulted in the data while the others did not. The problem is that we do not know which one is true: we simply have insufficient information to make a truth claim.

- We are thus inevitably led to formulate a *probability for different hypotheses, given the data* $p(\mathcal{H}\,|\,\mathcal{D})$. Deductive and inductive reasoning can therefore be summarised as resulting in two very different probabilities. For competing hypotheses $\mathcal{H}_m, m = 1, 2, 3\ldots$ and the same data $\mathcal{D}$

$$\begin{array}{rcll} \text{Deductive reasoning} & \Rightarrow & \text{probability of data given hypothesis} & = \quad p(\mathcal{D}\,|\,\mathcal{H}_m) \\ \text{Inductive reasoning} & \Rightarrow & \text{probability of hypothesis given data} & = \quad p(\mathcal{H}_m\,|\,\mathcal{D}) \end{array}$$

  In the Bayesian approach, both are needed and used. In the frequentist approach, only the deductive probability $p(\mathcal{D}\,|\,\mathcal{H})$ is permitted and used.

- We have introduced the solidus sign | as a necessary way of keeping track of which quantities are known and which are unknown/predicted within probabilities. For the purposes of calculating a probability, anything appearing to the right of a solidus is known and certain, while quantities to the left are unknown and uncertain.

- The fact that $\mathcal{H}_m$ appears once as a known and once as an unknown quantity may seem confusing. Indeed, speaking of "the probability of a truth $\mathcal{T}$" seems self-contradictory. The chain of

reasoning therefore also forces us to revise our language. We may no longer use words such as "truth" or "actual true law" in our reasoning and calculations but must instead always speak of one or more *hypotheses*. A hypothesis may or may not be true, and calculating the probability of it being true is logically consistent.

- Even if a particular hypothesis is considered true, the values of parameters entering the exact or approximate theory would often be unknown. In the case of parameters, the entire above arguments can be repeated on the level of different values for parameters θ; we would again have parameter-deductive probabilities $p(\mathcal{D} \,|\, \theta, \mathcal{H}_m)$ and parameter-inductive ones $p(\theta \,|\, \mathcal{D}, \mathcal{H}_m)$, both assuming $\mathcal{H}_m$ to be correct.

- Physicists often struggle to accept the change in mindset to speak about laws of nature as hypotheses. After all, they say, the basic laws of have been verified and tested many times, so that competing theories or hypotheses would therefore seem to be superfluous and one can safely speak of *the laws of nature* rather than *the most likely hypothesis for the laws of nature*. Indeed the big frameworks of laws as taught are powerful and universal, so that the probability that they are true is high. There can never, however, be total certainty. Most physicists agree that even the Standard Model, which has survived decades of experimental testing, is provisional and needs improvement.

  Other natural sciences recognise the need for competing hypotheses more easily because the number of variables involved is often very large while the amount and quality of data is not. It is easier for them to see that competing laws or explanations are possible.

- While the power of the ruling physics theories is undisputed, things are less clear-cut when the phenomenology (data and circumstances) is complicated and/or approximations to the exact theories have to be made. Experimentalists routinely use their data to infer parameter values and do curve fitting without perhaps realising that this is not an afterthought or technical detail. Theorists routinely plot the deductive consequences of their theories on top of given data without realising the dire need for a formal understanding of the inductive issues involved.

1.2 Examples

We briefly return to the examples provided to parse the deductive and inductive issues involved.

- *Flying intercontinental nuclear ballistic missile*
  - Variables: speed, trajectory, location of first detection, heat emission, radioactivity ...
  - Competing hypotheses: computer malfunction, detector malfunction, other flying objects, atmospheric distortions, decoys, ...
  - Deduction: work out possible values for the speed, trajectory, location etc for each hypothesis
  - Induction: based on available satellite and radio-dish data, work out the probability that this data originated from an actual ICBM.

- *Crime scene*
  - Variables: glass broken, suspicious persons, alarm signal, missing items, ...
  - Competing hypotheses (not mutually exclusive, which complicates things): actual break-in, cricket players in park, visiting relatives lost in suburbia, alarm system malfunction, ...

- *Elementary particle decay*
  - Variables: momenta and energies of decay particles, momenta and energies of other particles also measured,

- Competing hypotheses: (1) the bump in the invariant mass spectrum reflects an actual elementary particle (actually many) with a particular rest mass; (2) the bump is merely a statistical fluctuation in the noise of all possible decay particle combinations (the *look elsewhere effect*).
- Deduction: work out processed data in the form of relativistically invariant rest masses of all possible pairs, plot on a mass spectrum
- Induction: apply Bayes' Theorem based on reasonable priors and likelihoods for the particle and background

1.3 Stochastic systems

- A *system* is the closed or isolated entity under consideration in a particular situation or investigation. It can be a very real physical entity or an abstract gedankenexperiment. In both cases, it is a simplified idealised picture or description which deliberately ignores most or all of the complications which inevitably arise in a real-world situation. The emphasis falls neither on actual apparatus, nor on a holistic "*everything is connected to everything*" viewpoint, nor on the complications and mysteries of measurement. All of that is sacrificed for the sake of simplicity and clarity. Such simplicity comes at a price: the answers and behaviour of a system may not be applied to the real world without careful checking that the underlying assumptions are being satisfied. Caveat emptor.

- A system can be deterministic, stochastic, or a combination of the two.

- We mostly associate *determinism* with the notion of exact predictability. If, for example, we apply a differential equation such as Newton's Second Law to two isolated bodies and supply their initial positions and velocities, the resulting trajectories are ideally known for all future times with infinite precision.[1]

  Corresponding to the picture of determinism as exact evolution in time, there is also stochastic evolution, meaning that the differential equation or algorithm includes one or more terms or factors which are random and thereby unpredictable. Such systems are normally called *stochastic processes*.

- Randomness and stochasticity are not, however, tied to the concept of time evolution; there are two simpler and more fundamental levels. The first involves *repetition* of a particular algorithm or experiment such that each trial yields a different answer or result. On an even deeper level, stochasticity will be interpreted as *information deficiency*: randomness is the mathematical consequence of ignorance. We shall in this course consider only the levels of repetition and time evolution.

1.4 Simulation and inference

- In the computational context, deductive work is called *simulation* while inductive work is called *computational inference*.

- **Simulation** is, of course, the virtual representation or modelling of what may or may not be a real process. As a deductive method, simulation starts with a known analytical law or an algorithm which is assumed to be true.

[1] This naive picture fails easily. The trajectories of three or more Newtonian bodies can fall into the realm of *classical chaos* and become unpredictable.

Examples of deterministic analytical equations include differential equations, discrete evolution equations and integral equations. Algorithmic examples include chaotic maps, cellular automata and of course the discretised versions of differential equations.

Examples of stochastic simulation include the Monte Carlo stochastic processes which will form part of this course. On the simpler level of repetition, stochastic simulation assumes the truth of some probability distribution and simply draws samples from that distribution.

In stochastic simulation, the law is known and fixed, while the resulting data can vary.

- **Inference** starts with a known set of data which is assumed to be true. The goal is to use that known data plus background information or knowledge to determine which law or algorithm most likely created that data.

  In stochastic inference, the data is known and fixed, while the cause or law is unknown and can vary.

- **Simulation-inference-verification**: The combination of both simulation and inference into a single computer program code has created a powerful tool for exploring and testing both the simulation and inference aspects. The simulation component will typically created data based on a law or algorithm known, of course, to the programmer who coded it. The inference component will then pretend to not know that law or algorithm but try to reconstruct it based only on the available data. If the reconstruction method is good and valid, the inference process will yield the original simulation rule.

  That may sound trivial and circular and therefore a waste of time. Not so. Very often, the data does allow for alternative explanations or rules, especially if it has some stochastic component. Even for deterministic systems, the data may be highly sensitive to parameter values or initial conditions as well as to numerical error. Secondly, it is often impossible to simulate or infer a system exactly and approximations have to be used. The simulation-inference-verification cycle then serves as a valuable tool to assess the respective approximation methods themselves.

- Because the simulation and inference parts of the verification cycle often make use of the same functions and methods, it is tempting to use the same variables and program functions for both. That is a grave mistake which inevitably results in confusion and wrong conclusions. Simulation and inference are fundamentally different and must be kept apart even in a combined computer code. **It is crucial to separate simulation from inference in computational code combining the two.**

Chapter 2

Diffusion and the random walk

Determinism and stochasticity are in general not tied to the size of the system but to the available information. However, in physics, determinism has traditionally been associated with *macroscopic* "classical" phenomena, while the stochastic aspects are seen as dominant in *microscopic* ones. Thermodynamics and statistical physics are the best known examples: the thermodynamic relations of macroscopic quantities are mostly deterministic and macroscopic, while the underlying microscopic picture of atoms and their statistical physics description relies on randomness and probability.

One of the most important systems in physics which exhibit the duality of macroscopic classical behaviour and microscopic stochastic behaviour is that of *diffusion* versus *Brownian motion*, which is a special case of the family of *random walks*. Both are suitable for simulation and for inference, and both form the foundation for extensions and generalisations. We shall often refer to them in this course.

2.1 Maths technology I: delta functions and theta functions

- Here as elsewhere, we introduce mathematical topics as they arise, in this case delta functions and theta functions.

- For continuous x, the *Dirac delta function* $\delta(x)$ is defined to be zero everywhere except at x=0 where it is infinite while integrating to 1. The Dirac delta function is not a function in the conventional sense but rather the limit of a sequence of functions. One possible sequence is $\delta(x) = \lim_{k\to\infty} \delta_k(x)$ where

$$\delta_k(x) = \frac{k}{\sqrt{2\pi}} e^{-k^2x^2/2}. \tag{2.1}$$

  Comparison to the usual gaussian

$$p(x \,|\, 0, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2} \tag{2.2}$$

  shows that $\delta_k(x)$ is a gaussian with $1/k = \sigma$ the *standard deviation* which quantifies the width of the gaussian. The sequence of functions $\{\delta_k(x), k = 1, 2, 3, \ldots\}$ therefore have widths which decrease to zero as $k \to \infty$.

- The Dirac delta function and all its precursors integrate to 1,

$$\int_{-\infty}^{\infty} dx\, \delta_k(x) = 1, \qquad \int_{-\infty}^{\infty} dx\, \delta(x) = 1 \tag{2.3}$$

and the Delta function has the property that, for any function $f(x)$, and any line interval $\mathcal{A} \subset \mathbb{R}$

$$\int_{\mathcal{A}} dx\, \delta(x-a)\, f(x) = \begin{cases} f(a) & \text{for all } a \in \mathcal{A} \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

- The *Heaviside theta function* is an integral of the Dirac delta function,

$$\Theta(z) = \int_{-\infty}^{z} \delta(x)\, dx = \begin{cases} 1 & \text{whenever } z \geq 0 \\ 0 & \text{otherwise;} \end{cases} \tag{2.5}$$

  it has the form of a step function at $z = 0$. The step can be placed at any point a by translating the argument, $\Theta(z-a)$.

- Theta functions can be combined in many ways, resulting for example in *window functions* or *indicator functions*; for example

$$U(x \,|\, A, B) = \Theta(B-x)\, \Theta(x-A), \; = \; \begin{cases} 1 & \text{whenever } A \leq x \leq B \\ 0 & \text{otherwise} \end{cases} \tag{2.6}$$

  has the form of a box of unit height ranging from $x = A$ to $x = B$.

- For integer x, the corresponding functions are the Kronecker delta and the discrete Heaviside and indicator functions. The Kronecker delta is defined for integers a, b as

$$\delta_{a,b} = \begin{cases} 1 & \text{whenever } a = b \\ 0 & \text{otherwise} \end{cases} \tag{2.7}$$

  To make the indices readable, Kronecker deltas may be used in the following three equivalent notations,

$$\delta_{a,b} = \delta(a-b) = \delta(a,b) \tag{2.8}$$

  Correspondingly, for integers x, z and a,

$$\Theta(z) = \sum_{x=-\infty}^{z} \delta_{x,0} = \begin{cases} 1 & \text{whenever } z \geq 0 \\ 0 & \text{otherwise,} \end{cases} \tag{2.9}$$

  while $\Theta(z-a)$ will be 1 for all $z \geq a$ and zero elsewhere. A discrete indicator function is also easily defined. There is a simple "algebra of Kronecker deltas" encompassing rules such as additivity, commutativity and scaling. For any integers a, b, c

$$\delta(a+c, b+c) = \delta(a,b) \qquad \forall\, c \tag{2.10}$$

$$\delta(a,b) = \delta(b,a) \tag{2.11}$$

$$\delta(\alpha a, \alpha b) = \delta(a,b) \qquad \forall\, \alpha \in \mathbb{R}, \alpha \neq 0 \tag{2.12}$$

  Contrast the last equation with that of the Dirac delta function, for which $\delta(\alpha x) = \delta(x)/|\alpha|$. Further properties are, for example, $\delta(a,b)\delta(b,c) = \delta(a,c)\delta(b,c) \neq \delta(a,c)$ and for any integer subset $\mathcal{A} \subset \mathbb{Z}$

$$\sum_{x \in \mathcal{A}} \delta(x,b) f(x) = \begin{cases} f(b) & \text{whenever } b \in \mathcal{A} \\ 0 & \text{otherwise.} \end{cases} \tag{2.13}$$

2.2 Diffusion

- Diffusion is the macroscopic description of a microscopic process. The physical context is that we have a solvent such as water or air, into which molecules of another type are introduced and tracked. While the number of diffusing molecules is much smaller than those of the surrounding solvent, there are still enough to warrant a description in terms of a density. This can be either a mass density or a number density; the descriptions are equivalent since each diffusing molecule has the same mass. We shall use number density.

- The density can be quantified only if there is some minimum volume over which it can be determined; these are called fluid cells. A fluid cell must be small enough to warrant a "continuous" description of a density at a point $\boldsymbol{x}$ but large enough to contain a sizeable number of diffusing molecules.

- Before developing the theory of diffusion, we briefly remark on the context. In diffusion, the solvent is macroscopically static, constituting the simplest possible case of fluid dynamics. In general, the solvent may itself be flowing while the diffusing molecules are carried along with the solvent. The diffusing agent is then called a *passive scalar* which is *advected* by the solvent.

  The fluid dynamics of the solvent itself can be simple or increasingly complex. The intermediate case of incompressible fluids at constant temperature is governed by the *Navier-Stokes equation* which is nonlinear in the fluid velocity field $\boldsymbol{v}(\boldsymbol{x}, t)$,

$$\frac{\partial \boldsymbol{v}}{\partial t} + \boldsymbol{v}\cdot\boldsymbol{\nabla}\boldsymbol{v} = -\boldsymbol{\nabla}p + \nu\,\boldsymbol{\nabla}^2\boldsymbol{v}$$
$$\boldsymbol{\nabla}\cdot\boldsymbol{v} = 0$$

  and leads to highly complex phenomena such as vortices, turbulence and scaling behaviour. Beyond that lie the even more complex fluid systems with varying density and temperature. Fluid dynamics plays a large role in atmospheric physics, oceanography, aviation and astrophysics, to name but a few.

- We return to diffusion. On the macroscopic level, the diffusing agent is described not as a collection of point particles but as a density ρ, which is a function both of space coordinate $\boldsymbol{x}$ and of time t and is thereby a *scalar field*.

- Seen as a deductive theory, diffusion originates from combining two laws (hypothesis, assumed truth). The first, conservation of the diffusing agent molecules (they can move but cannot be created or destroyed) is captured in the continuity equation

$$\frac{\partial \rho(\boldsymbol{x}, t)}{\partial t} = -\boldsymbol{\nabla}\cdot\boldsymbol{j}(\boldsymbol{x}, t). \tag{2.14}$$

  stating that the increase in time of $\rho(\boldsymbol{x}, t)$ is the result of the net number of particles flowing into the fluid cell via a current density of moving molecules $\boldsymbol{j}(\boldsymbol{x}, t)$. The second law is Fick's First law, which states that the current density is proportional to the gradient in the concentration,

$$\boldsymbol{j}(\boldsymbol{x}, t) = -D\boldsymbol{\nabla}\rho(\boldsymbol{x}, t) \tag{2.15}$$

  where the proportionality constant $D > 0$ is called the *diffusion constant* or *diffusivity*. The continuity equation and Fick's law together yield the diffusion equation

$$\frac{\partial \rho(\boldsymbol{x}, t)}{\partial t} = D\,\boldsymbol{\nabla}^2\rho(\boldsymbol{x}, t) \tag{2.16}$$

- The partial differential equation (2.14) can be solved analytically; the solution is the gaussian distribution for the number density

$$\rho(\boldsymbol{x},t) = \frac{N}{(4\pi Dt)^{3/2}} e^{-\boldsymbol{x}^2/4Dt} \tag{2.17}$$

with $N > 0$ the total number of diffusing particles.

- If we consider only diffusion in one space dimension, the coordinate $\boldsymbol{x}$ becomes just x, the gradient $\boldsymbol{\nabla}$ becomes $(\partial/\partial x)$ and the diffusion equation becomes

$$\frac{\partial \rho(x,t)}{\partial t} = D\,\frac{\partial^2 \rho(x,t)}{\partial t^2} \tag{2.18}$$

with solution

$$\rho(x,t) = \frac{N}{(4\pi Dt)^{1/2}}\, e^{-x^2/4Dt} \tag{2.19}$$

- The correctness of this solution is easily verified by substituting it into the diffusion equation (2.18). It is that special solution which results from the assumption that all N of the diffusing agent molecules are at time $t = 0$ concentrated at the point $\boldsymbol{x} = 0$ with zero volume, meaning that the density at $t = 0$ is infinite at $\boldsymbol{x} = 0$ and zero elsewhere. As shown in Section 2.1, the Dirac delta function can be written as the limit of a sequence of distributions, $\delta(x) = \lim_{k\to\infty} \delta_k(x)$ with δ_k the gaussian of (2.15) whose width $1/k$ decreases to zero as $k \to \infty$. If in Eq. (2.19) we identify $k^2 = 1/(2Dt)$, it is clear that

$$\lim_{t\to 0} \rho(x,t) = N\,\delta(x). \tag{2.20}$$

2.3 Scale and dimension in diffusion

- We make use of the diffusion example to illustrate the role played by physical dimension, scales and dimensionless variables. This is important both for the resulting analytical insight as well as for simulations. While this is often overlooked, simulations on the computer are always conducted in dimensionless variables. Computers store dimensionless numbers, not Joules or metres or any other physical quantity.

- We first consider the physical dimensions (length $\mathcal{L}$ and time $\mathcal{T}$) of the various quantities for the case of one-dimensional diffusion, Eqs. (2.18)–(2.19). Clearly

$$\dim(\rho) = \mathcal{L}^{-1} \qquad \dim(t) = \mathcal{T} \qquad \dim(x) = \mathcal{L}$$

and since the dimensions on both side of Eqs. (2.18)–(2.19) must be the same, we deduce that

$$\dim(D) = \mathcal{L}^2\,\mathcal{T}^{-1} \tag{2.21}$$

$$\dim(\boldsymbol{j}) = \dim(D)\,\dim(x^{-1})\,\dim(\rho) = (\mathcal{L}^2\,\mathcal{T}^{-1})(\mathcal{L}^{-1})(\mathcal{L}^{-1}) = \mathcal{T}^{-1} \tag{2.22}$$

For diffusion in three dimensions, the only changes are

$$\dim(\rho) = \mathcal{L}^{-3} \qquad \dim(\boldsymbol{j}) = \mathcal{L}^{-2}\mathcal{T}^{-1}$$

- Staying with the example of diffusion in one space dimension, we now consider within this example three ways in which the physial dimensions and dimensionless variables can be used to make both the physics and mathematics more transparent. The one-dimensional diffusion solution of Eq. (2.19)

$$\rho(x,t) = \frac{N}{\sqrt{4\pi Dt}}\, e^{-x^2/4Dt} \tag{2.23}$$

will be written in various ways below.

- **A more appropriate system of physical units**
  - First we briefly survey the physical scales involved in diffusion. In the original Brownian motion, the diffusing agent was a pollen particle which was observed under the microscope to be moving randomly on the water surface. This suffices to estimate that the typical microscopic time scale must have been about $T_0 = 10^{-1}$ seconds, while the typical microscopic length scale would be larger than the size of an atom ($L_0 = 10^{-10}$ metres) but much smaller than the view field of the microscope; it turns out to be about 10^{-6} metres. A typical size for the diffusion coefficient would therefore be $D_0 = (10^{-6})^2(2\times 10^{-1})^{-1} = 5\times 10^{-12}\mathrm{m}^2\mathrm{s}^{-1}$. The reason for including the factor 2 will become apparent below.
  - Define a dimensionless space variable and a dimensionless time variable

$$\tau = \frac{t}{T_0} \qquad\qquad \xi = \frac{x}{L_0} \tag{2.24}$$

  - Rewrite the exponent and prefactor of Eq. (2.23) in terms of these dimensionless variables as well as a dimensionless diffusion constant $\delta = D/D_0$,

$$\frac{x^2}{4Dt} = \frac{\xi^2 L_0^2}{4D\tau T_0} = \frac{\xi^2}{2\tau\delta} \qquad\qquad 4\pi Dt = 2\pi\tau\delta L_0^2 \tag{2.25}$$

    and so

$$\rho(x,t) = \frac{N e^{-\xi^2/2\tau\delta}}{L_0\sqrt{2\pi\tau}} \tag{2.26}$$

    We identify $N/L_0 = \rho_0$ as an average density i.e. the total number of agents divided by the appropriate length scale. The diffusion solution can hence be written in terms of a dimensionless density,

$$\frac{\rho(x,t)}{\rho_0} = \frac{e^{-\xi^2/2\tau\delta}}{\sqrt{2\pi\tau\delta}} \tag{2.27}$$

    Comparison with this to the dimensioned solution (2.23) shows strong similarity, as it should. The difference lies not in a different mathematical form but in the fact that x, t, D will normally be rather small numbers, while their dimensionless counterparts ξ, τ, δ will be of order one, and the numerical implementation will therefore be more stable.[1]

- **Using physical scales to create a units system**
  - The choices for space and time scales above were helpful but arbitrary in the sense that they reflected only scale estimates in round decimals. It is normally a better idea to let the given scale constants define the units themselves rather than imposing some estimates.
    In the present example, the diffusion constant D is given by the properties of the solvent and diffusing agent. Rather than defining D in terms of T_0 and L_0, we hence invert the dependencies. We can choose either the pair (D, L_0) as independent or the pair (D, T_0).
  - The choice of (D, T_0) yields a dependent length scale $L_0 = \sqrt{2DT_0}$; hence

$$\tau = \frac{t}{T_0} \qquad\qquad \xi = \frac{x}{L_0} = \frac{x}{\sqrt{2DT_0}} \tag{2.28}$$

$$\frac{x^2}{4Dt} = \frac{\xi^2 2DT_0}{4Dt} = \frac{\xi^2}{2\tau} \qquad\qquad 4\pi Dt = 4\pi DT_0\tau = 2\pi\tau L_0^2 \tag{2.29}$$

[1] This may not seem important for the physical scales of this particular problem, but for more complicated multivariate problems, the dimensioned quantities can quickly become very small or large, exceeding the capabilities of floating-point or long data types.

resulting in

$$\frac{\rho(x,t)}{\rho_0} = \frac{e^{-\xi^2/2\tau}}{\sqrt{2\pi\tau}} \tag{2.30}$$

which closely resembles the previous case except that the δ has been rendered redundant by using D as a scale rather than the arbitrary D_0.

– The choice of (D, L_0) as independent scales yields a dependent time scale $T_0 = L_0^2/2D$; the factor 2 is not necessary but helpful in understanding later. Then the dimensionless variables, exponent and prefactor are

$$\tau = \frac{t}{T_0} = \frac{2Dt}{L_0^2} \qquad \xi = \frac{x}{L_0} \tag{2.31}$$

$$\frac{x^2}{4Dt} = \frac{\xi^2 L_0^2}{2\tau L_0^2} = \frac{\xi^2}{2\tau} \qquad 4\pi Dt = 4\pi D(\tau L_0^2/2D) = 2\pi\tau L_0^2 \tag{2.32}$$

and the same form for $\rho(x,t)/\rho_0$ as above.

- **Using observation variables**

– We can also use these techniques for macroscopic measurements or observations, choosing D plus a fixed observation time t to define the two independent scales, or D plus a fixed observation point x.

– Choosing (D, t) yields a dependent length scale $X(t) = \sqrt{2Dt}$; this amounts to taking a "spatial snapshot" of the density at fixed t. In this case

$$\tau = \frac{t}{t} = 1 \qquad \xi = \frac{x}{X(t)} = \frac{x}{\sqrt{2Dt}} \tag{2.33}$$

$$\frac{x^2}{4Dt} = \frac{x^2}{2X(t)^2} = \frac{\xi^2}{2} \qquad 4\pi Dt = 2\pi X(t)^2 \tag{2.34}$$

which on substitution result in

$$\rho(x,t) = \frac{Ne^{-x^2/2X(t)^2}}{X(t)\sqrt{2\pi}} = \frac{N}{X(t)}\frac{e^{-\xi^2/2}}{\sqrt{2\pi}} \tag{2.35}$$

The first form in terms of x shows how the "scale" $X(t)$ is simply the standard deviation σ of the gaussian. This also shows that the scale of the resulting spatial gaussian

$$X(t) = \sqrt{2Dt} \tag{2.36}$$

grows with the square root of time. This property will be encountered many more times.

– The choice (D, x) as independent scales yields a dependent time scale $T(x) = x^2/2D$; this is the case of "watching a movie" of $\rho(x,t)$ at fixed x as a function of time. The resulting variables are

$$\tau = \frac{t}{T(x)} = \frac{2Dt}{x^2} \qquad \xi = \frac{x}{x} = 1 \tag{2.37}$$

$$\frac{x^2}{4Dt} = \frac{1}{2\tau} = \frac{T(x)}{2t} \qquad 4\pi Dt = \frac{2\pi t x^2}{T(x)} \tag{2.38}$$

and so

$$\rho(x,t) = \frac{Ne^{-T(x)/2t}}{|x|\sqrt{2\pi t/T(x)}} = \frac{N}{|x|}\frac{e^{-1/2\tau}}{\sqrt{2\pi\tau}} \tag{2.39}$$

which converges to zero in the limit $\tau \to 0$ as it should.

2.4 Scale and dimension in general

We generalise the scale- and dimension-related concepts and issues raised by the above diffusion example.

(a) **Physical dimension**

- There are two classes of scales: those defined by the system or by nature itself and human-defined ones.
- Physical units defined by human convention such as the SI system are essentially arbitrary. When there are no other scales in a given system, they may be used as such. Whenever a system's natural scales deviate significantly from those human-defined ones, it is better to use those inherent in the system.
- By physical dimension we mean, in the case of nonrelativistic physics, the physical measurables time $\mathcal{T}$, mass $\mathcal{M}$ and distance or length $\mathcal{L}$, which constitute the basic framework of our physical world. They are, of course, crucial when you want to put in numbers, but they are also helpful in analysing the different terms and factors in mathematical equations.
- A particular type of law or governing equation may link two or more physical dimensions by a *constant of nature*, for example the speed of light c in relativity and the unit quantum $\hbar$ in quantum mechanics. Such constants must, of course, be included explicitly in any formulation using human-defined scales. While this is always correct, it is wise and elegant to use such constants to eliminate one of the relevant physical dimensions in favour of the other. Since, for example, $\dim(c) = \mathcal{L}\mathcal{T}^{-1}$, all times can and should be written in terms of $\mathcal{L}^{-1}$; alternatively, lengths can be written in terms of inverse time. When this is done, the symbol c automatically falls away.

  Note that electromagnetism is a particularly hard case because some conventions include factors 4π into the human-defined systems.
- We demonstrated in the diffusion example how two or more physical constants can be combined to give scales in other physical dimensions.

(b) **Scales:** A fundamental question which a physicist must ask of a given physical system is *What sets the physical scale or scales of the system?*

- *Scaling* involves physical systems which have no intrinsic physical scale. These special systems require special treatment and are not part of this course.
- We have already shown by example in diffusion how scales are set and can be used.
- Where a given system has more than one scale for a given physical dimension or combination of physical dimensions, it is always sensible to consider the resulting *scale ratios*. Normally, but not always, a situation where one scale is much larger than the other is simpler than where the two relevant scales have a similar magnitude.

  When the physical system is kind enough to us to carry widely differing scales, we speak of *scale separation*. Such scale separation often permits the formulation of a simplified theory in terms of effective variables. A typical example is the separation of atomic scales from those encountered in classical physics, which permits us to reformulate the problem as one of classical physics quantities.

(c) **Dimensional analysis**

- The basic principle of dimensional analysis is that *every term in a mathematical equation must have exactly the same powers of $\mathcal{M}$, of $\mathcal{T}$ and of $\mathcal{L}$.*

- A second principle of physical dimensions is that *The expression inside any trigonometric function, square root, exponent or logarithm must be dimensionless.* This, too, can be very useful in checking the correctness of equations and calculating with them.
- Probability densities: Dimensional analysis also immediately shows why there is a fundamental difference between probabilities of discrete x versus probabilities of continuous x. For discrete x, $p(x,t)$ must be dimensionless, but for continuous x the normalisation condition $\int dx\, p(x,t) = 1$ shows that $p(x,t)$ is a *probability density function* (PDF), with dimensions of density $\mathcal{L}^{-1}$ if x has dimension $\mathcal{L}$.

(d) **Dimensionless variables**

- Dimensionless variables are easily defined by dividing any given dimensioned variable by the appropriate physical scale. This is, as stated, necessary for any numerical implementation anyway. It results in a big advantage in that the numerical values of such variables will typically range around 1 rather than very large or very small numbers. For example, the physical mass of the electron has a scale of around 10^{-31} kg in SI units, and it would be very foolish to write a computer program for electrons in such units.
- Dimensionless variables have one disadvantage compared to dimensioned ones: it is no longer possible to check a long derivation by comparing the physical dimensions of individual terms in the equations.
- Essentially any equation which cannot be written in terms of polynomials must be reduced to dimensionless variables. It is, however, sometimes possible to multiply an overall dimensionless equation on both sides by factors which carry physical dimension.

(e) **Simulation** should ideally follow the steps set out above. Given a particular physical system or, these include:

- Identify the physical dimensions of the system
- Consider which constants of nature (such as $c, \hbar$) or constants of the problem (such as D in diffusion) are pertinent.
- Where appropriate, use the available constants as scales of that physical dimension, or else combine such constants into a quantity which has the physical dimension desired.
- Define dimensionless variables by dividing by the appropriate scale.
- Run the simulation in dimensionless variables
- Once results have been obtained, express these in terms of dimensioned physical quantities by multiplying the dimensionless results by the appropriate scale.

In a case where the simulation does not purport to represent or mimick any real physical system, the analysis in terms of physical dimension falls away. Nevertheless, it may be useful to analyse the algorithm or mathematical equation in terms of the scales of the numbers involved, and to divide variables by such scales in order to simplify the numerics.

(f) **Location:** Besides scale, the other fundamental question to ask about any physical system or problem is that of location. In general, one may define location as some measure of the point where the relevant probability is large. Measures of location may include the maximum (mode) the first expectation value (mean) and the median. For Gaussian probabilities, the location is usually defined in terms of the parameter μ, which is the mode, expectation value and median all in one. We shall not pursue the topic further here.

2.5 The simple random walk

The second of our standard examples is the so-called Simple Random Walk (SWR), the microscopic counterpart of the diffusion problem.

- Physically, the concept of a random walk probably originated in the discovery of Brownian motion as already discussed.
- Brownian motion forms the subject of papers written by Einstein from 1905 to 1908 which contributed significantly to establishing that matter was indeed composed of atoms and molecules. In parallel, Marian Smoluchowski in 1906 published a slightly different version of the theory.
- Both the Einstein and Smoluchowski derivations were couched in the language of statistical physics, and random walks are therefore closely associated with statistical physics. They can, however, also be understood purely as a stochastic process without reference to the physics and assumptions of thermodynamics and statistical mechanics.
- While the random walk retains its large role in physics as such, it has acquired an even larger role in modern computation as it forms the heart of many stochastic numerical algorithms, including the Markov Chain Monte Carlo (MCMC) method which will be covered later in this course.
- The original physical interpretation of a random walk naturally implied walks in the three dimensions of physical space (or two, in the case of motion on the surface of a liquid). The general mathematics of random walks can, however, be formulated and used in an arbitrary number of dimensions. Modern applications routinely make use of very high-dimensional spaces. We shall study the simplest version which is clearly the one-dimensional case.
- The Simple Random Walk we now introduce is simple in that only the bare bones of the method are retained. This has the advantage of mathematical tractability, and many analytical results can be obtained. It also forms the baseline and starting point for more difficult and realistic physical problems which can usually only be investigated numerically.
- The very simplest one-dimensional random walk consists of a single *walker* or *agent* or *particle*, which moves on the one-dimensional lattice of integers. Both the time and the space coordinate are dimensionless, and we shall use the symbols t and x with this understanding. If and where a real walker would move on a real lattice, the length and time coordinates can easily be made dimensionless by division by the appropriate unit of time and the lattice constant.
- The simple random walk can be formulated in terms of either *space positions* x_t or in terms of *steps* ε_t. In terms of space positions, the algorithm is as follows:
  - Start at time $t = 0$ at position x_0, or more generally posit a probability of starting positions $p(x_0)$.
  - At time $t = 1$, the agent jumps up or down by one unit to $x_0 \pm 1$. The probability for an up jump to $x_1 = x_0 - 1$ is $1 - \rho$ and for a down jump to $x_1 = x_0 + 1$ is ρ, where $0 \leqslant \rho \leqslant 1$. Often, we will use the notation $\alpha = 1-\rho, \beta = \rho, \alpha+\beta = 1$.
  - The stochastic process is repeated for all successive times; the agent jumps up from the previous position x_{t-1} by one unit at time t with probability α and down from x_{t-1} with probability β.
  - The walk ends after T steps and x_T is the final position.

  In terms of the step variable, the algorithm is:
  - Start at time $t = 0$ at position x_0, or more generally posit a probability of starting positions $p(x_0)$.

  - At time $t = 1$, the agent takes a step up or down by one unit $\varepsilon_1 = \pm 1$. The probability for an up step ($\varepsilon_1 = -1$) is α and for a down step ($\varepsilon_1 = +1$) is β.
  - The stochastic process is repeated for all successive times; at any time t the agent steps up with probability α and down with probability β.
  - The walk ends after T steps and $x_T = x_0 + \sum_{t=1}^{T} \varepsilon_t$ is the final position.

- Both algorithms are elegantly summarised in terms of Kronecker deltas as follows. The step-based algorithm is based on independence of all steps and the simple per-step probability.

$$p(\varepsilon_T, \varepsilon_{T-1}, \ldots, \varepsilon_2, \varepsilon_1) = \prod_{t=1}^{T} p(\varepsilon_t) \tag{2.40}$$

$$p(\varepsilon_t) = \alpha\, \delta(\varepsilon_t, -1) + \beta\, \delta(\varepsilon_t, +1) \qquad \forall\, t = 1, \ldots, T \tag{2.41}$$

  while the position-based algorithm using conditional probabilities uses the Markov property and the single-jump probability

$$p(x_T, x_{T-1}, \ldots, x_1 \,|\, x_0) = \prod_{t=1}^{T} p(x_t \,|\, x_{t-1}) \tag{2.42}$$

$$p(x_t \,|\, x_{t-1}) = \alpha\, \delta(x_t, x_{t-1} - 1) + \beta\, \delta(x_t, x_{t-1} + 1) \tag{2.43}$$

- It is easy to see that the two sets of variables are related. For all t,

$$\varepsilon_t = x_t - x_{t-1} \tag{2.44}$$

$$x_t = x_0 + \sum_{t'=1}^{t} x_{t'} \tag{2.45}$$

- We note that the random walk is a deductive algorithm just as diffusion is. The former is stochastic and the latter deterministic; the former is microscopic and the latter macroscopic.

- These probabilities and related issues will be explored in various ways below. We shall also show how the simple random walk converges to diffusion for large numbers of steps.

Chapter 3

Basic probability theory

In Chapter 1, we touched on the basic hierarchy with one law, rule or concept at the top and many specific cases, facts and data at the bottom. We also briefly introduced the critical role played by information as distinct from underlying physical reality on the one hand and the observer and his probabilities and calculations on the other. We briefly mentioned that uncertainty, i.e. insufficient information, provides a sufficient explanation of what is often called a *random variable*, which should therefore preferably be called *uncertainty variable* or just *variable*.

The uncertainty variable can be best understood as a *question* which has a number of possible answers. In terms of notation, we now use the upper-case letter X to denote such a variable while the lower-case symbol x denotes one of several possible answers. The set of all possible answers, as enumerated by the observer based on his information, is called the *sampling space* $\mathcal{A}(X) = \{x_1, x_2, \ldots\}$. Coupled with each possible answer, the observer will determine a *probability* $P(X{=}x)$, also based on the available information.

3.1 Axiomatic theory of probability

The three quantities $X, \mathcal{A}$ and $P(X)$ together define what is commonly called *probability theory*. We postpone most of the interpretation regarding their interpretation to the Honours course, because that discussion raises fundamental philosophical issues which, while hugely important, are not the focus of the present course. We content ourselves with observing that the *mathematics of probability* applies both to deductive simulation and to inductive inference. For the sampling probability introduced below, it also applies both to the frequentist and Bayesian framework. The mathematics set out in this and the next chapters is the same, while the use and interpretation may differ radically.

As in all mathematical theories, we are tasked to set out its *axioms*, those quantities and relations which are postulated to be true without further motivation. In the case of the frequentist viewpoint, the relations are captured in the *Kolmogorov Axioms*; the Bayesian framework is based on the *Cox axioms*. Before that, we recast the triad $X, \mathcal{A}(X), P(X)$ as purely mathematical quantities rather than data-driven ones.

Axiomatic probability theory is only properly defined if each of the following **three elements** are fully specified:

$$\begin{aligned} X &= \text{question or variable} \\ \mathcal{A}(X) &= \text{sampling space} \\ P(X) &= \text{probability} \end{aligned}$$

We consider each in turn.

3.1.1 The question or variable X

- X is a **question**, an **operator** or an **action** whose answer or outcome is uncertain.

- Not much more need be said about X at this stage except to emphasise that data, experiments and indeed rational knowledge can be acquired only once X is properly defined and, where necessary, made answerable by means of the appropriate experimental apparatus. It is also of great importance to consider beforehand which variable or physical quantity is to be measured. The mathematics itself does not supply the meaning and interpretation; that must be supplied by the user.

- While we limit ourselves initially to just one variable X, the framework is readily expanded to two or more variables $X_1, X_2, \ldots$ which may or may not have the same character (question, possible answers etc).

3.1.2 Sampling space $\mathcal{A}(X)$

- We use the notation $\mathcal{A}(X)$ or $\mathcal{A}_X$ or just $\mathcal{A}$ to denote the sampling space, which is the set of all possible elementary outcomes of X or answers to X. The sampling space may contain a countable or uncountable number of elements.

  Discrete outcomes do not necessarily have to be numbers; they can be any set of discrete properties. Often it is, however, useful to assign to each such outcome a natural number $0, 1, 2, \ldots$; for example

  $$\begin{array}{lcl} \text{Answer 1} = \text{“red”} & \longrightarrow & a_1 = 0 \\ \text{Answer 2} = \text{“green”} & \longrightarrow & a_2 = 1 \end{array}$$

- As in the case of formulating the question X, the task of listing possible answers and thus the sampling space is the responsibility of the practitioner. The process of doing so is inductive, meaning that the practitioner must translate whatever reality he or she wants to capture into probability theory by using the background knowledge and the available information. The mathematics, if any, can only formalise what the model builder puts into it. Likewise, the details of any mappings from real-world answers such as the red-green pair above to numbers depend entirely on the problem and the practitioner.

- If they are discrete numbers, we often write the outcomes as $\mathcal{A}(X) = \{a_1, a_2, \ldots, a_M\} = \{a_e\}_{e=1}^{M}$. For continuous $\mathcal{A}(X)$, we can often write it in terms of set theory, for example a line interval

  $$\mathcal{A}(X) = [a_{\min}, a_{\max}] = \{x \,|\, a_{\min} \leq x < a_{\max}\}$$

- Because $\mathcal{A}(X)$ in all cases is a set, probability theory likes to make use of set theory.

- There are *elementary* and *compound* outcomes and sampling spaces. Elementary outcomes are by definition “mutually exclusive”, i.e. any single answer or outcome of X is elementary in the sense that the question X is formulated in such a way that it is impossible to get two overlapping answers a_e and $a_{e'}$ at the same time. Equivalently, elementarity means that if a particular answer a_e was found to be true, all other possible answers are necessariliy false.

- A simple example is the number of eyes appearing on the face of a cubic die. The background knowledge and information is *I have a cubic die; it has six faces marked as usual with eyes.* The question would be $X =$ “*If I toss the die, which of the six faces will land facing up?*” The sampling space is $\mathcal{A}(X) = \{1, 2, 3, 4, 5, 6\}$. No particular face landing on top can be confused with any other, and so these outcomes are elementary.

- The background knowledge provided and/or the information may be such that more than one question may be asked. For the example of the die, we may ask an earlier question such as E = *Is the number appearing on top an even number?*. In this case, the sampling space would be $\mathcal{A}(E) = \{\text{Yes,No}\}$. Based on the answer, a *conditional* question could be asked such as $(X \,|\, E{=}\text{Yes})$ = *In the case that the number is known to be even, what possible number could it be?* The corresponding sample space is then $\mathcal{A}(X \,|\, E{=}\text{Yes}) = \{2, 4, 6\}$, while $\mathcal{A}(X \,|\, E{=}\text{No}) = \{1, 3, 5\}$. The background knowledge implicitly being used is that *The numbers* $2, 4, 6, \ldots$ *are even.*

- An example of an earlier question with more than two possible answers uses the floor function: F = *What is the value of* $\lfloor X/2 \rfloor$*?* which has $\mathcal{A}(F) = \{0, 1, 2, 3\}$ and $\mathcal{A}(X \,|\, F{=}0) = \{1\}$, $\mathcal{A}(X \,|\, F{=}1) = \{2, 3\}$, $\mathcal{A}(X \,|\, F{=}2) = \{4, 5\}$, $\mathcal{A}(X \,|\, F{=}3) = \{6\}$.

- These examples motivate the introduction of *compound outcomes* which group the elementary outcomes based on the answer to an earlier question. In terms of set theory, the sampling spaces of compound variables are *subsets* of the elementary sampling space, $\mathcal{A}_b \subset \mathcal{A}(X)$ where $b = 1, 2, 3, \ldots, B$ enumerates the earlier answers. The question then changes from the simpler (X = *Which elementary answer* a_e *do I get?*) to the hierarchical structure ($X \,|\, E$ = *Given the answer to* E*, which elementary answer to* X *is possible?*).

- **Partitions:** Initially, we prefer to work with nonintersecting subsets whose secondary answers are mutually exclusive, which is compactly expressed as $\mathcal{A}_b \cap \mathcal{A}_c = \emptyset \;\forall\, b \neq c$ or more loosely $\mathcal{A}_b \cap \mathcal{A}_c = \delta_{a,b}\mathcal{A}_b$. A partition of $\mathcal{A}$ is hence defined as the set of subsets $\mathcal{A}_b$ for which

$$\mathcal{A}_b \cap \mathcal{A}_c = \delta_{bc}\,\mathcal{A}_b \quad \forall b, c = 1, 2, \ldots, B \tag{3.1}$$

$$\cup_{b=1}^{B}\,\mathcal{A}_b = \mathcal{A} \tag{3.2}$$

- The language of set theory is closely connected to the symbols of logic as follows:

Symbol	Explanation	Logical interpretation
$\mathcal{A}$	total sampling space	always TRUE
$\emptyset$	empty set	always FALSE
$\mathcal{A}_c$, $\mathcal{A}_b$ etc	possible subsets of $\mathcal{A}$	Statement "$\mathcal{A}_b$ is true" for the particular outcome
$\cap$	intersection of two sets	AND
$\cup$	union of two sets	OR
$\bar{A}_b$	complement of $\mathcal{A}_b$, i.e. all elements of $\mathcal{A}$ that do NOT belong to $\mathcal{A}_b$	NOT

3.1.3 Probability $P(X)$

The third component of a probability theory is the probability itself.

- In the frequentist mindset, which we do not support and consider inferior, probability is based on the so-called Kolmogorov axioms as follows.

Given a random variable X with sampling space $\mathcal{A}(X)$ and any subset $\mathcal{A}_b \subset \mathcal{A}$, there exists a real function $P(X)$ with the following axiomatic properties:

Kolmogorov I:	$P(X{\in}\mathcal{A}_b) \geq 0$
Kolmogorov II:	$P(X{\in}\mathcal{A}) = 1$
Kolmogorov III:	For mutually exclusive subsets $\mathcal{A}_1$ and $\mathcal{A}_2$
	$P(X{\in}\mathcal{A}_1{\cup}\mathcal{A}_2) = P(X{\in}\mathcal{A}_1) + P(X{\in}\mathcal{A}_2)$

- **Theorems:** The Kolmogorov axioms suffice to prove the following theorems fully. $\mathcal{A}_b$ and $\mathcal{A}_c$ are any two subsets of $\mathcal{A}$ and may intersect.

  1–1 $\mathcal{A}_b \subset \mathcal{A}_c \quad \Rightarrow \quad P(X{\in}\mathcal{A}_b) \leq P(X{\in}\mathcal{A}_c)$

  1–2 $0 \leq P(X{\in}\mathcal{A}_b) \leq 1 \quad \forall \mathcal{A}_b \subseteq \mathcal{A}$

  1–3 $P(\emptyset) = 0$

  1–4 $P(X{\in}\bar{\mathcal{A}}_b) = 1 - P(X{\in}\mathcal{A}_b)$ where $\bar{\mathcal{A}}_b$ is the complement of $\mathcal{A}_b$.

  1–5 If $\{\mathcal{A}_b\}_{b=1}^{B}$ is a partition of $\mathcal{A}$, then $\quad \sum_{b=1}^{B} P(X{\in}\mathcal{A}_b) = 1$

  1–6 When $\mathcal{A}_b$ and $\mathcal{A}_c$ have a nonempty intersection, the correct relation is

  $$P(X{\in}\mathcal{A}_b{\cup}\mathcal{A}_c) = P(X{\in}\mathcal{A}_b) + P(X{\in}\mathcal{A}_c) - P(X{\in}\mathcal{A}_b{\cap}\mathcal{A}_c).$$

  1–7 For any subsets $\mathcal{A}_a$ and $\mathcal{A}_b$ it is true that

  $$P(X{\in}\mathcal{A}_b) = P(X{\in}\mathcal{A}_b{\cap}\mathcal{A}_a) + P(X{\in}\mathcal{A}_b{\cap}\bar{\mathcal{A}}_a)$$

  1–8 For any partition $\{\mathcal{A}_b\}$ of $\mathcal{A}$ and any other variable Y defining a subset of the same overall $\mathcal{A}$, $\quad P(Y) = \sum_b P(Y \cap \mathcal{A}_b)$

- In the Bayesian framework, probability is based on degree of belief in the truth of logical propositions as follows.

  Notation: Let the degree of belief in proposition A = "the question X has answer a" be denoted by $\mathcal{B}(A)$. The degree of belief in a conditional proposition "A, given information $\mathcal{I}$ and assuming hypothesis $\mathcal{H}$ to be true" is represented by $\mathcal{B}(A\,|\,\mathcal{H},\mathcal{I})$. For the sake of clarity, we omit $\mathcal{H}$ from the notation below.

  Cox Axiom 1: *Degrees of belief can be* ***ordered****: if $\mathcal{B}(A\,|\,\mathcal{I})$ is greater than $\mathcal{B}(A'\,|\,\mathcal{I})$, and if $\mathcal{B}(A'\,|\,\mathcal{I})$ is greater than $\mathcal{B}(A''\,|\,\mathcal{I})$, then $\mathcal{B}(A\,|\,\mathcal{I})$ is greater than $\mathcal{B}(A''\,|\,\mathcal{I})$.*

  Cox Axiom 2: *The degrees of belief in a proposition A and in its negation $\bar{A}$ are* ***related****; there is a function f such that*

  $$\mathcal{B}(A\,|\,\mathcal{I}) = f[\mathcal{B}(\bar{A}\,|\,\mathcal{I})].$$

  Cox Axiom 3: *The degree of belief in a* ***conjunction of propositions*** *A and A' is related to the degree of belief in the conditional proposition $A\,|\,A'$ and the degree of belief in the proposition A'. There is thus a function g such that*

  $$\mathcal{B}(A \text{ and } A'\,|\,\mathcal{I}) = g\left[\mathcal{B}(A\,|\,A',\mathcal{I}), \mathcal{B}(A'\,|\,\mathcal{I})\right]$$

- It is not the purpose of this course to explore the philosophical foundations of the above two sets of axioms; we shall merely take note and consider the practical consequences.

- The Cox Axioms together with the known rules of Boolean Algebra are sufficient to reproduce the entire set of theorems shown above. Couched in terms of set theory, there is therefore no difference between the mathematical consequences of using Kolmogorov or Cox axioms.

- There is, however, a huge difference in the interpretation and scope or applicability of this mathematics. The frequentist world view permits only deductive probability, while in the Bayesian view both deductive and inductive probability is permitted and needed.

- Specifically, the Cox Axiom 2 leads directly to the probability *sum rule* while the Cox Axiom 3 leads to the probability *product rule* of any propositions A and A' which, it turns out, play just the role of the variables or questions discussed so far. Given the available information $\mathcal{I}$, degrees of belief result in quantifiable probabilities obeying

$$P(A\,|\,\mathcal{I}) = 1 - P(\bar{A}\,|\,\mathcal{I}) \tag{3.3}$$
$$P(A \text{ AND } A'\,|\,\mathcal{I}) = P(A\,|\,A',\mathcal{I})\,P(A'\,|\,\mathcal{I}), \tag{3.4}$$

  where, for example $P(A\,|\,\mathcal{I})$ is the statement *The probability that logical proposition A is true, given $\mathcal{I}$* and so on. From these two rules and the axioms, all other relations follow, for example also the generalised OR probability of Theorem 1–6,

$$P(A \text{ OR } A') = P(A) + P(A') - P(A \text{ AND } A'). \tag{3.5}$$

  For more than one logical statement whose answers are mutually exclusive, we also again obtain what is colloquially called normalisation. If $\{A_b\}_{b=1}^{B}$ is a set of propositions which are mutually exclusive in the sense that $P(A_b \text{ AND } A_c\,|\,\mathcal{I}) = 0$ for all $b \neq c$, then as usual

$$\sum_{b=1}^{B} P(A_b\,|\,\mathcal{I}) = 1. \tag{3.6}$$

- Importantly, in the Bayesian framework all probabilities are *conditional probabilities*, if only to always make explicit that they depend on the available information. Naturally, whenever there are competing hypotheses and thereby different probabilities, the appropriate $\mathcal{H}_m$ symbols must also be included. It is also clear that the difference to the set-theoretic theorems is only one of language used.

- **Notation:** While the most general formulation of probability is that of logical statements, we will increasingly use the generic symbol X as representing the variable, and $\mathcal{A}(X)$ as the set of possible answers or outcomes to X. By default, outcomes are assumed to be elementary, so for a finite set of M outcomes $\mathcal{A}(X) = \{a_e\}_{e=1}^{M}$. The probability for X having a particular answer a_e can then be written in a proper but cumbersome uppercase-P-notation, or a compact but ambiguous lowecase-p-notation as

$$\begin{aligned} P(X{=}x\,|\,\mathcal{I}) &= p(x) \quad && \text{for } \textit{any } x \in \{a_1,\ldots,a_M\} \\ P(X{=}a_e\,|\,\mathcal{I}) &= p(a_e) && \text{when a } \textit{particular} \text{ outcome is considered} \end{aligned} \tag{3.7}$$

  To simplify notation, we often omit the $\mathcal{I}$ symbol from the notation.

3.1.4 Principle of Indifference

In general, it is not easy to assign probabilities, but in the case of ignorance we are aided by the *Principle of Indifference*. If the available information $\mathcal{I}_0$ only tells us that there are M possible outcomes $\{a_e\}_{e=1}^{M}$ but that there is no other feature or property to distinguish between them. The Principle of Indifference states that, given $\mathcal{I}_0$, the observer must assign equal probability to all,

$$p(a_e\,|\,\mathcal{I}_0) = p(a_{e'}\,|\,\mathcal{I}_0) \quad \forall\, e, e' = 1,\ldots,M. \tag{3.8}$$

Since as stated all elementary outcomes are by assumption distinguishable and exhaustive, the probabilities sum to unity,

$$\sum_{e=1}^{M_0} p(a_e \,|\, \mathcal{I}_0) = 1$$

and since all $p(a_e \,|\, \mathcal{I}_0)$ are equal, we obtain

$$\boxed{p(a_e \,|\, \mathcal{I}_0) = \frac{1}{M} \qquad \forall e} \tag{3.9}$$

The Principle of Indifference is also known as the *Principle of Insufficient Reason*.

3.2 Multivariate probability: generally valid properties

- **Variable with two properties:** If a variable has outcomes which consist of two properties tested by X_1 and X_2 respectively, it is written as a list (X_1, X_2). The comma implies the AND logic, i.e. $(X_1, X_2) \equiv (X_1 \text{ AND } X_2)$. Two examples are (a) repetition of the same question or variable and (b) the probability for three-momentum consists of three different answers, namely the components p_x, p_y and p_z; therefore $P(\boldsymbol{p}) = P(p_x, p_y, p_z)$.

- Correspondingly, the triad of random variable, sample space and probability must be generalised to $(X_1, X_2), \mathcal{A}(X_1, X_2), P(X_1, X_2)$ or in vector form $\boldsymbol{X}, \mathcal{A}(\boldsymbol{X}), P(\boldsymbol{X})$.

- In general, there is no assumption of time ordering between X_1 and X_2. If such ordering exists, it must be explicitly specified.

- The sampling space $\mathcal{A}(X_1, X_2)$ has similarities and difference with regard to the one-variable $\mathcal{A}(X)$ considered so far. They are similar in that both are sets of possible answers or outcomes, and both should be complete in the sense of containing all of them. Of course, $\mathcal{A}(X_1, X_2)$ contains all possible *joint* answers such as *the ball is red AND it carries the label 2* or *the elementary particle has a mass m AND a spin s.* This illustrates that $\mathcal{A}(X_1, X_2)$ may be more complicated, both because the variables X_1 and X_2 may differ in character and because of possible dependencies of one answer on the other. In mathematical terms, such dependencies would manifest themselves as *nonfactorising sample spaces*,

  $$\mathcal{A}(X_1, X_2) \neq \mathcal{A}(X_1) \otimes \mathcal{A}(X_2)$$

  where $\mathcal{A}(X_1)$ would contain all possible answers to X_1 and likewise $\mathcal{A}(X_2)$ all the answers to X_2. It is, for example, not true that the joint sampling space for elementary particles $\mathcal{A}(X_1{=}\text{mass}, X_2{=}\text{spin})$ would contain particles with all possible masses and all possible spins.

- Multivariate probabilities can be projected onto **marginal distributions** of fewer variables. Let's say we have joint sampling space $\mathcal{A}(X_1, X_2) = \{(a_e, c_b) \,|\, e{=}1, 2, \ldots M_1, b = 1, 2, \ldots, M_2\}$, where some of the possible joint answers may have zero probability. Marginalisation means asking for only one of the available joint answers or outcomes (say $X_1{=}a_e$), regardless of the answer to X_2. Then the *marginal probability* for this X_1-answer is

  $$P(X_1{=}a_e) = \sum_{x_2 \in \mathcal{A}(X_2 \,|\, X_1 = a_e)} P(X_1{=}a_e, X_2{=}x_2{=}c_b). \tag{3.10}$$

  The *conditional sampling space* $\mathcal{A}(X_2 \,|\, X_1{=}a_e)$ means that the sum over x_2 runs over all elements $X_2{=}c_b$ of $\mathcal{A}(X_1, X_2)$ for which X_1 is fixed to a_e. This conditional space can change in dependence

on a_e, so it is important to keep track of it. We will often simplify this cumbersome notation to read

$$p_1(x_1) = \sum_{x_2 \in \mathcal{A}(X_2 \,|\, x_1)} p(x_1, x_2) \tag{3.11}$$

where x_1 would be one of the outcomes a_e and x_2 one of the outcomes c_b. It is crucial to understand that (3.11) is not one equation but M_1 different equations, one for each possible $X_1{=}x_1{=}a_e$.

- Projection onto $X_2{=}c_b$ has the same structure and results from the same line of thought, so we merely record that

$$p_2(x_2) = \sum_{x_1 \in \mathcal{A}(X_1 \,|\, x_2)} p(x_1, x_2) \tag{3.12}$$

- We have used subscripted notation p_1 and p_2 only to emphasise that the resulting functions often differ from each other.

- The marginals themselves are automatically properly normalised,

$$1 = \sum_{x_1 \in \mathcal{A}(X_1)} p_1(x_1) \qquad\qquad 1 = \sum_{x_2 \in \mathcal{A}(X_2)} p_2(x_2) \tag{3.13}$$

where $\mathcal{A}(X_1)$ is the projection of $\mathcal{A}(X_1, X_2)$ onto X_1 and similarly for $\mathcal{A}(X_2)$.

- The term *marginalisation* is helpful if we visualise the joint probabilities of discrete outcomes (X_1, X_2) as a table where each row represents the probability of an outcome of X_1 and each column the probability of an outcome of X_2. The values of the marginal probability $p(X_1)$ then can be written as a "column vector" along the vertical margin of the matrix; likewise the marginal $p(X_2)$ can be written as a "row vector" along the bottom margin.

  For example, let the sampling space be $\mathcal{A}(X_1, X_2) = \{(0,4), (0,5), (0,6), (1,4), (1,5), (1,6)\}$. The corresponding 2×3 table of joint probabilities $p(X_1, X_2)$ is shown below,[1] along with the corresponding marginals. The normalisation of the joint probability for this example implies

$$a + b + c + d + e + f = 1.$$

		$X_2 =$			
		4	5	6	$P(X_1)$
		$P(X_1, X_2) =$			
$X_1 =$	0	a	b	c	$a{+}b{+}c$
	1	d	e	f	$d{+}e{+}f$
$P(X_2)$		$a{+}d$	$b{+}e$	$c{+}f$	

- As discussed, **conditional probabilities** are built into the foundations of the theory, so that

$$p(x_2 \,|\, x_1) \;=\; \frac{p(x_1, x_2)}{p_1(x_1)} \;=\; \frac{p(x_1, x_2)}{\sum_{x_2} p(x_1, x_2)} \tag{3.14}$$

$$p(x_1 \,|\, x_2) \;=\; \frac{p(x_1, x_2)}{p_2(x_2)} \;=\; \frac{p(x_1, x_2)}{\sum_{x_1} p(x_1, x_2)} \tag{3.15}$$

[1] The table technique is useful for two random variables but obviously fails if we have three or more variables's.

are just re-arrangements of the product rule. There is a separate conditional probability $p(x_2 \,|\, x_1)$ for each value for $X_1{=}x_1$, and the $p(x_2 \,|\, X_1)$ can differ for different values of X_1. The sampling spaces for the marginals in the denominators are those of Eq. (3.11) and (3.12).

- Continuing the above example in the table below, all the values for the conditional probabilities $p(X_1 \,|\, X_2)$ are shown in the left-hand table and for $p(X_2 \,|\, X_1)$ in the right-hand one.

	$X_2 =$		
	4	5	6
	$p(x_1 \,\|\, x_2) =$		
$X_1 = 0$	$\frac{a}{a+d}$	$\frac{b}{b+e}$	$\frac{c}{c+f}$
1	$\frac{d}{a+d}$	$\frac{e}{b+e}$	$\frac{f}{c+f}$
$\sum_{x_1} p(x_1 \,\|\, x_2)$	1	1	1

	$X_2 =$			
	4	5	6	$\sum_{x_2} p(x_2 \,\|\, x_1)$
	$p(x_2 \,\|\, x_1) =$			
$X_1 = 0$	$\frac{a}{a+b+c}$	$\frac{b}{a+b+c}$	$\frac{c}{a+b+c}$	1
1	$\frac{d}{d+e+f}$	$\frac{e}{d+e+f}$	$\frac{f}{d+e+f}$	1

- The example also reminds us that normalisation applies only to the variable left of the vertical line:

$$\sum_{x_1} p(x_1 \,|\, x_2) \;=\; 1 \qquad \forall\, x_2 \tag{3.16}$$

$$\sum_{x_2} p(x_2 \,|\, x_1) \;=\; 1 \qquad \forall\, x_1 \tag{3.17}$$

3.3 Independence

We now turn to the property of independence which is used quite commonly but is not generally true; on the contrary, it is an assumption. The resulting factorisation is of course very convenient, but convenience does not imply truth. Independence comes in three different degrees: statistical, logical and conditional independence.

- In the frequentist viewpoint, so-called *statistical independence* (SI) of variables X_1 and X_2 is motivated by proof or hypothesis of the physical independence of the respective outcomes. In that case, the joint probability factorises into the product of its marginals

$$p(x_1, x_2) \overset{\text{SI}}{=} p(x_1)\, p(x_2) \qquad \forall\, (x_1, x_2) \in \mathcal{A}(X_1, X_2),$$

  and we say that "X is statistically independent of X_2". A well-known example is the ideal gas in statistical physics in which it is reasonable to assume that the approximation of N noninteracting single molecules is good enough for many calculations. The relation $p(x_1 \,|\, x_2) \overset{\text{SI}}{=} p(X_1)$ follows directly via the product rule $p(x_1, x_2) = p(x_1 \,|\, x_2)\, p(x_2)$ and can be used as an alternative definition.

- Bayesian statistics includes the more general concept of *logical independence* which is based on the available information: There may or may not be physical independence, but we have no information regarding dependence. In the absence of such explicit information, we choose to work with a hypothesis that there is no dependence rather than introduce an arbitrary version of dependence. Logical independence therefore is not necessarily a statement of physical independence of the events or properties which are described by X_1 and X_2, and it is a weaker condition than statistical independence. The mathematics of LI is, however, identical with that of SI, namely

$$p(x_1 \,|\, x_2) \overset{\text{LI}}{=} p_1(x_1) \qquad \forall (x_1, x_2) \in \mathcal{A}(X_1, X_2) \tag{3.18}$$

or equivalently $p(x_1, x_2) \overset{\text{LI}}{=} p_1(x_1)\, p_2(x_2)$.

- Neither SI nor LI follow from the axioms but is an additional assumption. To emphasise that fact, we put the letters "LI" over the equal sign.

- The third type of independence requires even fewer assumptions and is therefore most widely applicable. Two variables X_1 and X_2 are conditionally independent if and only if a specific third variable or proposition C is true, i.e.

$$P(X_1 \,|\, X_2, C) \overset{\text{CI}}{=} P(X_1 \,|\, C) \tag{3.19}$$

$$\text{and also} \qquad P(X_1 \,|\, X_2) \neq P(X_1) \text{ in general.} \tag{3.20}$$

  The information contained in C therefore makes knowledge of X_2 superfluous while X_2 may still remain relevant to X_1 otherwise.

- Example continued: The first table below represents the original joint probability $p(x_1, x_2)$ of our above example, while the second table lists the products $p(x_1)p(x_2)$. Simple comparison of the values shows that X_1 and X_2 are **not** statistically independent in this case.

		$X_2 =$			
		4	5	6	$p(x_1)$
		$p(x_1, x_2) =$			
$X_1 =$	0	a	b	c	$a{+}b{+}c$
	1	d	e	f	$d{+}e{+}f$
$p(x_2)$		$a{+}d$	$b{+}e$	$c{+}f$	

		$X_2 =$			
		4	5	6	$p(x_1)$
		$p(x_1)p(x_2) =$			
$X_1 =$	0	$(a{+}b{+}c)(a{+}d)$	$(a{+}b{+}c)(b{+}e)$	$(a{+}b{+}c)(c{+}f)$	$a{+}b{+}c$
	1	$(d{+}e{+}f)(a{+}d)$	$(d{+}e{+}f)(b{+}e)$	$(d{+}e{+}f)(c{+}f)$	$d{+}e{+}f$
$p(x_2)$		$a{+}d$	$b{+}e$	$c{+}f$	

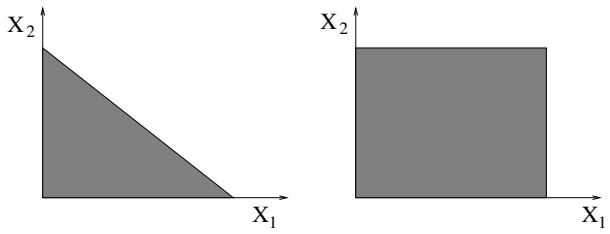
- **Independence and sample spaces:** The factorisation of probabilities (3.18) defining statistical independence implies that the joint sample space must also factorise into the sample spaces of the marginals,

$$P(X_1, X_2) \overset{\text{LI}}{=} P(X_1)\, P(X_2) \qquad \Rightarrow \qquad \mathcal{A}(X_1, X_2) = \mathcal{A}(X_1) \otimes \mathcal{A}(X_2) \tag{3.21}$$

  where $\otimes$ is the outer product. A nonfactorising $\mathcal{A}(X_1, X_2)$ necessarily means that X_1 and X_2 are not independent. By contrast a factorising sample space does not necessarily imply independence,

$$\mathcal{A}(X_1, X_2) = \mathcal{A}(X_1) \otimes \mathcal{A}(X_2) \qquad \not\Rightarrow \qquad P(X_1, X_2) = P(X_1)\, P(X_2)$$

  If for example $P(X_1, X_2)$ is uniform on the nonfactorising triangular $\mathcal{A}(X_1, X_2)$ shown on the left below, the outcome space for $P(X_1)\, P(X_2)$ is the product $\mathcal{A}(X_1) \otimes \mathcal{A}(X_2)$ shown on the right. Neither the marginal probabilities nor their product are uniform.



- We have already seen in the case of the simple random walk that the walk positions are not independent; only the individual steps are. The factorisation of the joint sampling space cannot therefore be assumed in general. Another case of nonfactorising sampling space involves the change of a parameter determining the stochastic behaviour; for example, if you change the voltage on your apparatus between X_1 and X_2, you probably must already reconsider whether the new measurement(s) have the same outcome space.

3.4 Repetition

We return to the general case, i.e. do not assume independence except in some examples.

- As already stated, two random variables X_1 and X_2 can either refer to a repetition of the same experiment, or two different aspects of the same experiment.
- Case *two different properties:* The sampling spaces of X_1 and X_2 will differ; they can, for example, have different physical dimensions such as length and energy. There is in this case no chance of confusing an outcome of X_1 with the corresponding outcome for X_2: the two parts of the joint outcome are distinguishable.
- In the case of *independent repetition of the same experiment*, the variables X_1 and X_2 and any further repetitions have the same sampling space, $\mathcal{A}(X_1) = \mathcal{A}(X_2) = \cdots$ which we can hence just call $\mathcal{A}$. By definition of independence, we also have

$$P(X_1, X_2, \ldots, X_N) = \prod_{i=1}^{N} P(X_i). \tag{3.22}$$

3.5 Occupation numbers

- *Inherent indistinguishability* arises for systems with discrete-valued sampling spaces where the same outcome a_e may arise for different variables X_i and the information on the order of specific outcomes is not available.
- If, for example, $\mathcal{A}(X) = \{3, 7\}$ so B=2 and we have two variables (X_1, X_2), so N=2 also. The possible outcome pairs are $\mathcal{A}(X_1, X_2) = \{(3,3), (3,7), (7,3), (7,7)\}$. On the level of (X_1, X_2), the outcomes $(3, 7)$ and $(7, 3)$ are distinct, but once the ordering information is lost, they map onto the same variable, namely the fact that there is one 3 and one 7 in the result. For $\mathcal{A}_{b=1} = \{3\}$ and $\mathcal{A}_{b=2} = \{7\}$, the corresponding occupation numbers for this case would be $(n_1, n_2) = (2,0), (1,1), (1,1), (0,2)$. Correspondingly, the probabilities are related by

$$\begin{aligned} P(N_1{=}2, N_2{=}0) &= P(X_1{=}3, X_2{=}3) \\ P(N_1{=}1, N_2{=}1) &= P(X_1{=}3, X_2{=}7) + P(X_1{=}7, X_2{=}3) \\ P(N_1{=}0, N_2{=}2) &= P(X_1{=}7, X_2{=}7) \end{aligned}$$

  and $\mathcal{A}(N_1, N_2) = \{(2,0), (1,1), (0,2)\}$ as well as $P(N_1, N_2)$ have just three possible outcomes.

- The change from individual outcomes to occupation numbers is a noninvertible **projection**

$$\boldsymbol{x} = \{x_i\}_{i=1}^{N} \qquad \longrightarrow \qquad \boldsymbol{n} = \{n_b\}_{b=1}^{B} \quad \text{with} \quad \sum_{b=1}^{B} n_b \stackrel{!}{=} N \tag{3.23}$$

- Mostly $B \ll N$, but the numbers N and B are independent and it is possible for B to be larger than N. Even in that case, the transformation of the variable, sampling space and probability

$$\begin{aligned} \boldsymbol{X} = (X_1, \ldots, X_n) \qquad &\longrightarrow \qquad (N_1, \ldots, N_B) = \boldsymbol{N} \\ \mathcal{A}(X_1, \ldots, X_n) \qquad &\longrightarrow \qquad \mathcal{A}(N_1, \ldots, N_B) \\ P(X_1, \ldots, X_n) \qquad &\longrightarrow \qquad P(N_1, \ldots, N_B). \end{aligned} \tag{3.24}$$

  may not be invertible.

- While it would be consistent to keep uppercase letters for the occupation number variables, we usually write lower-case letters $\mathcal{A}(n_1, \ldots, n_B)$ and $P(n_1, \ldots, n_B)$.

- One consequence of projection is that independence is lost:

$$\begin{aligned} P(X_1, X_2, \ldots, X_N) &\stackrel{\text{LI}}{=} P(X_1)P(X_2)\cdots P(X_N) \\ \text{but} \qquad P(n_1, n_2, \ldots, n_B) &\neq P(n_1)\,P(n_2)\cdots P(n_B) \end{aligned}$$

  This is trivially apparent from the fact that the B occupation number variables must always add up as $\sum_{b=1}^{N} n_b = N$, so that at least one of them cannot be specified independently.

- The general mathematics of finding the projection from $P(X_1, \ldots, X_N)$ to $P(N_1, \ldots, N_B)$ is beyond the scope of this course. We state only that the probability for the occupation numbers, $P(\boldsymbol{N})$ is found by listing all outcomes $\boldsymbol{x}$ for the variables $\boldsymbol{X}$, matching each one with a corresponding outcome $\boldsymbol{n}$, and then adding up all the probabilities $P(\boldsymbol{X}{=}\boldsymbol{x})$ whose $\boldsymbol{x}$ corresponds to a specific $\boldsymbol{n}$ to find that $P(\boldsymbol{N}{=}\boldsymbol{n})$. Some details will be provided in Sections 4.1 and 4.2.

- By contrast, it is fairly easy to specify the sampling space for occupation numbers; it is

$$\mathcal{A}(\boldsymbol{n}) = \{\boldsymbol{n} \,|\, (0 \leqslant n_b \leqslant N)\ \forall\, b = 1, \ldots, B; \quad \textstyle\sum_{b=1}^{B} n_b \stackrel{!}{=} N\} \tag{3.25}$$

  For the simplest binomial case $B = 2$, this simplifies to

$$\begin{aligned} \mathcal{A}(n_1, n_2) &= \{(n_1, n_2) \,|\, n_1 = 0, 1, 2, \ldots, N \quad \text{with} \quad n_2 = N{-}n_1\} \\ &= \{(N, 0), (N{-}1, 1), \ldots, (1, N{-}1), (0, N)\} \end{aligned}$$

- For $B < N$, the projection from $\{X_i\}$ to $\{n_b\}$ usually results in a huge reduction in the sampling space. For example, given $|\mathcal{A}(X_i)| = 2$ for a single coin toss, the size of the sample space for 30 *distinguishable* tosses is $|\mathcal{A}(X_1, \ldots, X_{30})|$ is $2^{30} \simeq 10^9$, while $|\mathcal{A}(n_{\text{Tails}}, n_{\text{Heads}})| = 31$ because $n_{\text{Heads}} \in \{0, 1, \ldots, 30\}$ and $n_{\text{Tails}} = 30 - n_{\text{Heads}}$.

- Evidently, the occupation number probabilities are necessarily **combinations** of elementary outcomes. *The Principle of Indifference cannot therefore be used directly on occupation numbers.* On the contrary, it is required to first pretend that the outcomes are elementary and to apply the *Principle of Indifference* on the elementary level, and afterwards to project onto occupation numbers.

3.6 Ordering, interval probabilities and continuous sampling spaces

- We return briefly to the axioms and theorems of Section 3 to set the scene. They are valid for a wide class of subsets of $\mathcal{A}(X)$, including sets made up of elementary outcomes and for sets consisting of many elementary outcomes. They are also true for intersecting subsets $A \cap B \neq \emptyset$ and for outcomes which have no particular ordering.

- Often, however, we have no need for this generality but can specialise to sampling spaces with properties that we commonly encounter in physics and other applications. Specifically, if the outcomes a_e of a discrete-valued $\mathcal{A}$ are both elementary and numerically ordered integers, then we can define "interval probabilities" for all outcomes between a minimum x_a and maximum x_b,

$$P(x_a \leq X \leq x_b) = \sum_{x_a \leq X \leq x_b} P(X) \tag{3.26}$$

  where it is assumed that the interval boundaries x_a and x_b are in $\mathcal{A}$ or on its boundary. Another convention uses the inequality in the upper bound, $x_a \leq X < x_b$.

- Interval probabilities can be easily extended to cover lattices with noninteger lattice constants. There is no fundamental limitation of $\mathcal{A}(X)$ to integers.

- Interval probabilities on a continuous sampling space can correspondingly be defined as

$$P(x_a \leq X \leq x_b) = \int_{x_a}^{x_b} P(X{=}x)\, dx \tag{3.27}$$

- In this language, it becomes clear that we have already implicitly assumed both the ordering and elementary character of outcomes x in continuous sampling spaces, since a statement such as $x > y$ is an elementary property of real numbers, and $x = y$ similarly implies that they are the *same* outcome to infinite precision if they represent the same real number in $\mathbb{R}$.

- The integral already represented a warning that probabilities in continuous spaces and in discrete spaces need to be handled differently, because the "bin width" of a real number x is essentially zero. Put differently, the probability of ever finding an outcome for X which is *exactly*, to infinite accuracy, equal to some real outcome x is zero, in agreement with the Principle of Indifference (3.9) with $M \to \infty$. For this reason, there is a need to define an infinitesimal "bin width" dx, so that the product $P(X{=}x)\, dx$ is nonzero.

  Probability in continuous sampling space is therefore necessarily a *probability density* with dimensions $\dim(p(x)) = \dim(dx)^{-1}$. We often talk about a "probability density function" or PDF. This transition from a dimensionless to a dimensioned definition of probability has many important consequences, and it is therefore necessary always to treat the cases of discrete X and continuous X separately. An immediate example is that the PDF $P(X{=}x)$ can easily be larger than 1 for continuous X since only $P(X{=}x)\, dx$ must be smaller than 1.

- Since the outcome usually is sufficiently descriptive in ordered sets, we can extend our shorthand notaion to both the discrete and continuous cases as follows:

$$p(a_e) = P(X{=}a_e) \qquad \text{for discrete } X, \tag{3.28}$$

$$p(x) = \lim_{dx \to 0} \frac{1}{dx} P(X \in [x, x{+}dx]) \qquad \text{for continuous } X. \tag{3.29}$$

  The symbol p (and correspondingly P) is thus overloaded; it denotes either probability or probability density. The implied meaning is usually clear from the argument x.

- We can now generalise to continous outcomes many of the points of the previous sections. Assuming that such continuous $\mathcal{A}(X)$ represents a compact finite or infinite set on $\mathbb{R}$ (or $\mathbb{R}^N$ for multivariate continous sampling spaces, we have

$$\int_{\mathcal{A}(X)} dx\, p(x) = 1 \qquad \text{normalisation, one variable} \tag{3.30}$$

$$\int_{\mathcal{A}(X_1,X_2)} dx_1\, dx_2\, p(x_1,x_2) = 1 \qquad \text{normalisation, joint probability} \tag{3.31}$$

$$\int_{\mathcal{A}(X_1,X_2)} dx_2\, p(x_1,x_2) = p_1(x_1) \qquad \text{marginal distribution} \tag{3.32}$$

$$\int_{\mathcal{A}(X_1,X_2)} dx_1\, p(x_1,x_2) = p_2(x_2) \qquad \text{marginal distribution} \tag{3.33}$$

$$p(x_1 \,|\, x_2) = \frac{p(x_1,x_2)}{p_2(x_2)} \qquad \text{conditional probability} \tag{3.34}$$

$$p(x_2 \,|\, x_1) = \frac{p(x_1,x_2)}{p_1(x_1)} \qquad \text{conditional probability} \tag{3.35}$$

$$p(x_1,x_2) = p(x_1 \,|\, x_2)\, p_2(x_2) = p(x_2 \,|\, x_1)\, p_1(x_1) \qquad \text{product rule} \tag{3.36}$$

$$\int_{\mathcal{A}(X_1)} dx_1\, p_1(x_1) = 1 \qquad \text{normalisation of marginal} \tag{3.37}$$

$$\int_{\mathcal{A}(X_2)} dx_2\, p_2(x_2) = 1 \qquad \text{normalisation of marginal} \tag{3.38}$$

Generally, we can see that the formulae for discrete and continuous cases are identical except that sums are replaced by integrals and dimensionless probabilities are replaced by probability density functions.

3.7 Cumulative probability

- The so-called *cumulative probability distribution function* or simply *distribution function* F is a special case of an interval probability. Abbreviated as CDF, it is defined as the sum of all probabilities of outcomes up to a given maximum outcome x,

$$F(X{=}x) = F(x) = \begin{cases} \displaystyle\sum_{x' \le x} p(x') & \text{for the discrete case,} \\ \displaystyle\int_{-\infty}^{x} dx'\, p(x') & \text{for the continuous case.} \end{cases} \tag{3.39}$$

The lower limit $-\infty$ in the latter case covers both the cases where the lower bound of $\mathcal{A}(X)$ is finite or $-\infty$. In both cases, it is clear that

$$F(x < x_{\min}) = 0 \qquad \text{where } x_{\min} \text{ is the lower bound of } \mathcal{A}, \tag{3.40}$$

$$F(x \ge x_{\max}) = 1 \qquad \text{where } x_{\max} \text{ is the upper bound of } \mathcal{A}. \tag{3.41}$$

For discrete X, F is a step function which jumps by an amount of exactly $p(a_e)$ at every outcome $X{=}a_e$,

$$F(a_e) - F(a_{e-1}) = p(a_e) \tag{3.42}$$

while for continuous X, the derivative is

$$\frac{dF}{dx} = p(x). \tag{3.43}$$

The derivative $p(x)$ does not always exist even when F does. For that reason, some books prefer to couch everything in terms of F. If the derivative does exist, the compact notation

$$dF = p(x)\,dx \tag{3.44}$$

is still useful.

- Since $p(x) \geq 0$ always, F is a nondecreasing function.
- As shown later, F plays a very important role in the generation of random numbers.
- Cumulative distribution functions play a central role in the field of *order statistics* which addresses questions such as the "probability of the largest of N variables having a value x".
- Example for continuous x: Exponential distribution

$$p(x\,|\,\alpha) = \alpha\, e^{-\alpha x} \qquad \mathcal{A}(X) = \{x\,|\,0 \leq x \leq \infty\} \qquad \alpha > 0. \tag{3.45}$$

$$F(x\,|\,\alpha) = \int_0^x dx'\,\alpha e^{-\alpha x'} = \left(\frac{-\alpha}{\alpha}\right)\left[e^{-\alpha x} - 1\right] = 1 - e^{-\alpha x} \tag{3.46}$$

$$F(0\,|\,\alpha) = 0 \qquad F(\infty\,|\,\alpha) = 1 \tag{3.47}$$

- Example for discrete x: Geometric distribution. With the help of $\sum_{n=0}^{\infty} \rho_n = 1/(1-\rho)$, we have

$$p(n\,|\,\rho) = (1-\rho)\,\rho^n \qquad \mathcal{A}(n) = \{0,1,2,\ldots\} = \mathbb{N}_0, \qquad 0 < \rho < 1 \tag{3.48}$$

$$F(n\,|\,\rho) = \sum_{m=0}^{n} p(m\,|\,\rho) = (1-\rho)\left(\frac{1-\rho^{n+1}}{1-\rho}\right) = 1 - \rho^{n+1}. \tag{3.49}$$

$$p(n\,|\,\rho) = F(n\,|\,\rho) - F(n-1\,|\,\rho) = (1-\rho^{n+1}) - (1-\rho^n) = (1-\rho)\,\rho^n. \tag{3.50}$$

$$F(-1) = 1 - \rho^{-1+1} = 0 \qquad F(\infty) = 1 - \rho^{\infty} = 1 \tag{3.51}$$

Since $F(n)$ is negative for $n \leqslant -2$, it may be wise to insert a discrete Heaviside function $\Theta(n)$ into $F(n)$ to prevent such errors.

Chapter 4

Transformations

4.1 Transformations for discrete sampling spaces

4.1.1 One variable

- Transformations are basic to many mathematical and physical situations. Let us first consider the discrete case, and suppose we have a transformation from X to a new variable U. It is then critical to transform also the sampling space and probability alongside with the variable itself,

$$X \to U = \Phi(X) \qquad \mathcal{A}(X) \to \mathcal{A}(U) \qquad P(X) \to P(U). \tag{4.1}$$

- In the case of **bijective transformations** $P(X) \to P(U)$, the principle is simple: *every elementary outcome keeps its particular probability.* For bijective transformations, it is therefore just a matter of re-assigning the original probability at X to the transformed point $U = \Phi(X)$.

- For example, let $X \in \mathcal{A}(X) = \{0, 1, 2\}$ with corresponding probability $P(X{=}0) = \frac{3}{6}$, $P(X{=}1) = \frac{2}{6}$ and $P(X{=}2) = \frac{1}{6}$. For $U = \Phi(X) = X^2$, we obtain $\mathcal{A}(U) = \{0, 1, 4\}$, $P(U{=}0) = \frac{3}{6}$, $P(U{=}1) = \frac{2}{6}$ and $P(U{=}4) = \frac{1}{6}$. To generalise, note that

$$\begin{aligned} P(U{=}0) = P(X^2{=}0) = P(X{=}\sqrt{0}) = P(X{=}0) = \tfrac{3}{6} \\ P(U{=}1) = P(X^2{=}1) = P(X{=}\sqrt{1}) = P(X{=}1) = \tfrac{2}{6} \\ P(U{=}4) = P(X^2{=}4) = P(X{=}\sqrt{4}) = P(X{=}2) = \tfrac{1}{6} \end{aligned}$$

  where we made explicit use of our information that $X \geq 0$ in choosing the positive branch of the square root.

- The general formula for bijective transformations is

$$U = \Phi(X) \qquad X = \Phi^{-1}(U) \qquad P(U{=}\Phi(X){=}u) \;=\; P(X{=}\Phi^{-1}(u)) \tag{4.2}$$

  or, if we introduce the notation $g(u) \equiv P(U{=}u)$, in shorthand

$$g(u) = p(x{=}\Phi^{-1}(u)). \tag{4.3}$$

  While the notation may look complicated, the matter is actually simple: the probabilities remain the same. All that changes as a result of the transformation is the sampling space.

- In the case of **nonbijective transformations**, things are more complicated. To understand the issues that arise for nonbijective transformations, again consider an example: Let $X \in \mathcal{A}(X) = \{-2, -1, 0, 1, 2\}$, $P(X) = \{\frac{2}{10}, \frac{4}{10}, \frac{1}{10}, \frac{2}{10}, \frac{1}{10}\}$, and $U = \Phi(X) = X^2$. We now have two outcomes

map onto one, e.g. $X{=}2$ and $X{=}-2 \to U{=}4$. According to our stated principle, this means that

$$\begin{aligned} P(U{=}0) &= P(X{=}\sqrt{0}) = \tfrac{1}{10} \\ P(U{=}1) &= P(X{=}-\sqrt{1}) + P(X{=}+\sqrt{1}) = \tfrac{4}{10} + \tfrac{2}{10} = \tfrac{6}{10} \\ P(U{=}4) &= P(X{=}-\sqrt{2}) + P(X{=}+\sqrt{2}) = \tfrac{2}{10} + \tfrac{1}{10} = \tfrac{3}{10} \end{aligned}$$

- The general case is neatly covered by the relation

$$P(U{=}u) = \sum_{x \in \mathcal{A}(X)} \delta(u, \Phi(x))\; P(X{=}x) \tag{4.4}$$

$$\text{i.e.} \quad g(u) = \sum_{x \in \mathcal{A}(X)} \delta(u, \Phi(x))\; p(x) \tag{4.5}$$

where the Kronecker delta ensures that only the appropriate outcomes are actually counted for a given u.

- This relation is a direct consequence of using our standard rules, once we recognise that the transformation can be written as a "probability" with one and only one element in its $\mathcal{A}(U)$ sampling space for each x, as captured in the Kronecker delta, $g(u \,|\, x) = \delta(u, \Phi(x))$, from which, using the marginalisation rule, we obtain

$$g(u) = \sum_{x \in \mathcal{A}(X)} p(u, x) = \sum_{x \in \mathcal{A}(X)} p(u \,|\, x)\, p(x) = \sum_{x \in \mathcal{A}(X)} \delta(u, \Phi(x))\, p(x) \tag{4.6}$$

- Sometimes the nonbijective transformation can be handled by splitting it up into a sequence of individually bijective ones and, after applying Eq. (4.3), summing the probabilities for the same u. This less rigorous but perhaps more intuitive process must always be measured against Eq. (4.4).

4.1.2 Two or more variables

- The generalisation to two or more variables is straightforward; we merely extend (4.1) to

$$U = \Phi_1(X, Y), \quad V = \Phi_2(X, Y) \tag{4.7}$$

with inverse functions

$$X = \Psi_1(U, V), \quad Y = \Psi_2(U, V) \tag{4.8}$$

$$\mathcal{A}(X, Y) \to \mathcal{A}(U, V), \tag{4.9}$$

$$P(X, Y) \to P(U, V). \tag{4.10}$$

- For bijective transformations, the transformation relation for the probabilities is

$$g(u, v) = p[x{=}\Psi_1(u, v)\,,\; y{=}\Psi_2(u, v)] \tag{4.11}$$

while the sampling spaces again transform on a point-by-point basis as in the one-variable case.

- An example may help. Let

$$\begin{aligned} \mathcal{A}(X, Y) &= \{1, 2, 3\} \otimes \{\tfrac{1}{2}, \tfrac{2}{3}\} = \{(1, \tfrac{1}{2}), \ldots, (3, \tfrac{2}{3})\}, \\ U = \Phi_1(X, Y) &= \sqrt{\tfrac{1}{2}(X + Y)} \qquad\qquad V = \Phi_2(X, Y) = \tfrac{1}{2}(X - Y) \\ X = \Psi_1(U, V) &= U^2 + V \qquad\qquad Y = \Psi_2(U, V) = U^2 - V \\ \mathcal{A}(U, V) &= \{(\sqrt{\tfrac{3}{4}}, \tfrac{1}{4}), \ldots (\sqrt{\tfrac{11}{6}}, \tfrac{7}{6})\} \end{aligned}$$

from which, for example,

$$P\left[(U,V)=\left(\sqrt{\tfrac{11}{6}},\tfrac{7}{6}\right)\right]=P\left[X=\left(\sqrt{\tfrac{11}{6}}\right)^2+\tfrac{7}{6}=3, Y=\left(\sqrt{\tfrac{11}{6}}\right)^2-\tfrac{7}{6}=\tfrac{2}{3}\right]$$

- **Independence can be made apparent or hidden by transformation**. The same probability can be either dependent or possibly independent, given a suitable transformation to new variates.

  Take for example the **sum and difference of distinguishable dice**. The sampling space is clearly the set of number pairs from $(1,1)$ to $(6,6)$ which we can write in matrix form although the sampling space is not a matrix at all,

$$\mathcal{A}(X_1,X_2)=\left\{\begin{array}{cccccc}(1,1)&(2,1)&(3,1)&(4,1)&(5,1)&(6,1)\\(1,2)&(2,2)&(3,2)&(4,2)&(5,2)&(6,2)\\(1,3)&(2,3)&(3,3)&(4,3)&(5,3)&(6,3)\\(1,4)&(2,4)&(3,4)&(4,4)&(5,4)&(6,4)\\(1,5)&(2,5)&(3,5)&(4,5)&(5,5)&(6,5)\\(1,6)&(2,6)&(3,6)&(4,6)&(5,6)&(6,6)\end{array}\right\}=\mathcal{A}(X_1)\otimes\mathcal{A}(X_2),$$

  and again applying the Principle of Indifference and Statistical Independence, the joint probability and the marginals are, in "matrix" form,

		X_2						
		1	2	3	4	5	6	$P(X_1)$
X_1	1	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
	2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
	3	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
	4	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
	5	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
	6	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{6}$
$P(X_2)$		$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	

  i.e. X_1 and X_2 are independent, $P(X_1,X_2)=P(X_1)\,P(X_2)$. If we transform (X_1,X_2) to the sum and difference,

$$U=X_1+X_2,\qquad V=X_1-X_2$$

  then we get for the joint outcome (U,V) an sampling space which we rotate to exhibit its symmetry

$$\mathcal{A}(U,V)=\left\{\begin{array}{ccccccccccc}&&&&&(+7,+5)&&&&&\\&&&&(+6,+4)&&(+8,+4)&&&&\\&&&(+5,+3)&&(+7,+3)&&(+9,+3)&&&\\&&(+4,+2)&&(+6,+2)&&(+8,+2)&&(+10,+2)&&\\&(+3,+1)&&(+5,+1)&&(+7,+1)&&(+9,+1)&&(+11,+1)&\\(+2,\ 0)&&(+4,\ 0)&&(+6,\ 0)&&(+8,\ 0)&&(+10,\ 0)&&(+12,\ 0)\\&(+3,-1)&&(+5,-1)&&(+7,-1)&&(+9,-1)&&(+11,-1)&\\&&(+4,-2)&&(+6,-2)&&(+8,-2)&&(+10,-2)&&\\&&&(+5,-3)&&(+7,-3)&&(+9,-3)&&&\\&&&&(+6,-4)&&(+8,-4)&&&&\\&&&&&(+7,-5)&&&&&\end{array}\right\}$$

and since the marginal sampling spaces are

$$\mathcal{A}(U) = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$
$$\mathcal{A}(V) = \{-5, -4, -3, -2, -1, 0, +1, +2, +3, +4, +5\}$$

the sampling space certainly does not factorise, $\mathcal{A}(U,V) \neq \mathcal{A}(U) \otimes \mathcal{A}(V)$. and the joint probability $P(U,V)$ clearly does not factorise into its marginals $P(U)$ and $P(V)$ either, as shown in the table below.

		U											
		+2	+3	+4	+5	+6	+7	+8	+9	+10	+11	+12	$P(V)$
	+5						$\frac{1}{36}$						$\frac{1}{36}$
	+4					$\frac{1}{36}$		$\frac{1}{36}$					$\frac{2}{36}$
	+3				$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$				$\frac{3}{36}$
	+2			$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$			$\frac{4}{36}$
	+1		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{5}{36}$
V	0	$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$	$\frac{6}{36}$
	-1		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{5}{36}$
	-2			$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$			$\frac{4}{36}$
	-3				$\frac{1}{36}$		$\frac{1}{36}$		$\frac{1}{36}$				$\frac{3}{36}$
	-4					$\frac{1}{36}$		$\frac{1}{36}$					$\frac{2}{36}$
	-5						$\frac{1}{36}$						$\frac{1}{36}$
	$P(U)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	

4.2 Transformation for continuous sampling spaces

4.2.1 One variable

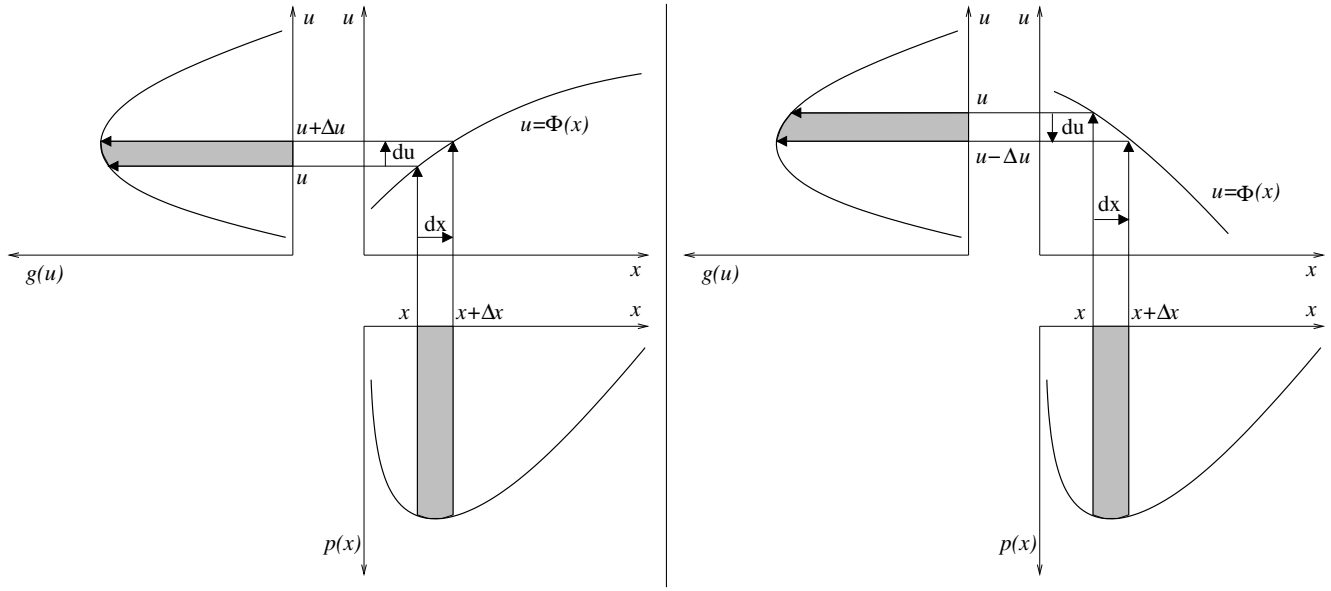


Figure 4.1: Transformation of continuous variables for positive slope (left) and negative slope (right)

- The situation for cases where $\mathcal{A}(X)$ is continuous is sketched in Figure 4.1. Axes have been rotated so as to align the respective x- and u-axes. The transformation $u = \Phi(x)$ is assumed to be known, differentiable and invertible. Both cases $du/dx > 0$ and $du/dx < 0$ are shown.

- The principle applied is that **probability, not probability density, must be conserved under transformation**. This implies that we must carefully distinguish between widths $\Delta x, \Delta u$, which by definition are always positive, and infinitesimal line elements dx, du, which can be positive or negative. All four are assumed to be infinitesimally small but nonzero. In differential form, the principle reads

$$\boxed{p(x)\,\Delta x \stackrel{!}{=} g(u)\,\Delta u.} \tag{4.12}$$

where $p(x)$ and $g(u)$ are *probability densities*. To explore the implications, we write this in terms of proper interval *probabilities*,

$$P(x \le X \le x + \Delta x) \stackrel{!}{=} \begin{cases} P(u \le U \le u + \Delta u) & \text{when } (du/dx) > 0, \\ P(u - \Delta u \le U \le u) & \text{when } (du/dx) < 0, \end{cases} \tag{4.13}$$

where $(u \pm \Delta u) = \Phi(x + \Delta x)$ are determined by the transformation. Rewriting the above as integrals, we must distinguish the two cases depending on the sign of $du = \pm\,\Delta u$. Depending on the sign of the transformation slope, we get two different mappings. For positive slope, the result is straightforward,

$$\text{If } \frac{du}{dx} > 0, \text{ then } \quad \int_x^{x+\Delta x} dx\, p(x) = \int_u^{u+\Delta u} du\, g(u) = \int_x^{x+\Delta x} dx\, g(u(x)) \left(\frac{du}{dx}\right), \tag{4.14}$$

while for negative slope we must use $du = -\Delta u$ to keep the area positive,

$$\text{If } \frac{du}{dx} < 0, \text{ then } \quad \int_x^{x+\Delta x} dx\, p(x) = \int_u^{u-\Delta u} (-du)\, g(u) = \int_x^{x+\Delta x} dx\, g(u(x)) \left(\frac{-du}{dx}\right). \tag{4.15}$$

The two cases can be consolidated into

$$\int_x^{x+\Delta x} dx\, p(x) = \int_x^{x+\Delta x} dx\, g(u(x)) \left|\frac{du}{dx}\right| \tag{4.16}$$

and since this equation is an identity for all x and all possible $p(x)$ and all possible $\Phi(x)$ the integrands must be identical. Hence

$$p(x) = g(u{=}\Phi(x)) \left|\frac{du}{dx}\right|$$

or, on inverting,

$$\boxed{g(u) = p(x{=}\Phi^{-1}(u)) \left|\frac{dx}{du}\right|} \tag{4.17}$$

where $p(x = \Phi^{-1}(u))$ on the right hand side implies that x should be substituted by the inverse transformation so that the result is written in terms of u only.

- In parallel to the discrete case (4.6), we can generalise this by use of our standard rules of probability on recognising that the transformation $u = \Phi(x)$ can be written as a "conditional probability" with only one possible outcome for each x, namely $p(u \,|\, x) = \delta(u - \Phi(x))$. On marginalising and using the product rule, we have immediately

$$g(u) = \int_{\mathcal{A}(X\,|\,u)} dx\, p(u, x) = \int_{\mathcal{A}(X\,|\,u)} dx\, p(u \,|\, x)\, p(x) = \int_{\mathcal{A}(X\,|\,u)} dx\, \delta(u - \Phi(x))\, p(x) \tag{4.18}$$

and the special case (4.17) is recovered by the usual rules of inverting a Dirac delta function from $\delta(u - \Phi(x))$ to one in terms of x.

- **Transformation of interval probabilities:** In contrast to the differential transformation law (4.17), we can write the transformation law *under the integral* without absolute values and quite generally. For any function $h(x)$, including of course trivial cases such as $h(x) = 1$, and for any subset $\mathcal{A}_b(X) \subset \mathcal{A}(X)$, the transformation identity is cast in terms of the image subset $\mathcal{A}_b(X) \longrightarrow \mathcal{A}_b(U) \subset \mathcal{A}(U)$ as

$$\int_{\mathcal{A}_b(U)} h(u{=}\Phi(x))\, g(u)\, du = \int_{\mathcal{A}_b(X)} h(x)\, p(x)\, dx \tag{4.19}$$

- The transformation of (cumulative) distribution functions is an important special case. For the CDF, $h(x) = 1$ and we limit ourselves to positive (du/dx). For positive slope, $\mathcal{A}(X') = (x_{\min}, x)$ transforms to $\mathcal{A}(U') = (u_{\min}, u{=}\Phi(x))$ and we obtain

$$F_X(x) = \int_{x_{\min}}^{x} dx'\, p(x') = \int_{u_{\min}}^{\Phi(x)} du'\, g(u') = F_U(u{=}\Phi(x)). \tag{4.20}$$

  The case of negative (du/dx) is rather ambiguous and so we do not consider it further here.

- It may happen that we know the initial and final probability densities $g(u)$ and $p(x)$ and wish to find the transformation $u = \Phi(x)$ which connects them. Eq. (4.20) immediately provides the answer: Given that $F_U(u)$ and $F_X(X)$ are functions of u and x respectively, we simply solve the align for u,

$$F_U(u) = F_X(x) \qquad \Rightarrow \qquad u = F_U^{-1}[F_X(x)] \equiv \Phi(x). \tag{4.21}$$

  This forms the basis for the generation of nonuniform random numbers as shown later in this course.

- Care must be taken in the **transformation of the sampling space** $\mathcal{A}(X) \to \mathcal{A}(U)$. If $\mathcal{A}(X)$ is some compact set with well-defined boundaries $x_{\min}$ and $x_{\max}$, a linear transformation $\Phi(x)$ will result in a similar compact set (line interval) between $\Phi(x_{\min})$ and $\Phi(x_{\max})$. If, however, Φ is nonlinear, we must take a good deal more trouble to find $\mathcal{A}(U)$.

- **Nonbijective transformations:** As in the case of discrete variables, such cases are handled by splitting up the nonbijective transformation into a sum of intervals or "branches" such that the transformation in each branch is bijective. The probability density $g(u)$ for any particular u is then the sum of all densities $p(u)$ that map onto that u times the absolute value of the derivative,

$$g(u) = \sum_b p(x(u))_b \left|\frac{dx}{du}\right|_b \tag{4.22}$$

  where b runs over all the branches whose x values map onto u.[1]

[1] One of the properties of the Dirac delta function $\delta(f(x))$ of a function $f(x)$ is that it can be written as a sum over terms taken at each and every point x_0 where $f(x_0) = 0$,

$$\int \delta(f(x))\, p(x)\, dx = \sum_{\{x_0 \,|\, f(x_0)=0\}} \frac{p(x_0)}{|f'(x_0)|}.$$

Taking $f(x) = \Phi(x) - u$ so that $f'(x) = d\Phi/dx = du/dx$, Eq. (4.18) can also be written as

$$g(u) = \sum_{\{x_0 \,|\, \Phi(x_0)=u\}} p(x_0) \left|\frac{dx}{du}\right|_{x_0} = \sum_{x \in \mathcal{A}(X)} p(x) \left|\frac{dx}{du}\right| \delta(u, \Phi(x))$$

where $\delta(u, \Phi(x))$ is the same Kronecker delta that appears in Eq. (4.5).

- Here is an example illustrating a nonbijective transformation. Let $\mathcal{A}(X) = \{x \,|\, -3 < x < 6\}$ and $p(x) = x^2/81$. Transforming to $U = X^2 = \Phi(X)$, we have $\mathcal{A}(U) = \{u \,|\, 0 \le u < 36\}$. Split $\Phi(X)$ into two branches over $\mathcal{A}_1(X) = \{-3 < x < 0\}$ and $\mathcal{A}_2(X) = \{0 \le x < 6\}$. The inverse transformation of the two branches are

$$\begin{aligned} \Psi_1(x) &= -\sqrt{u} \qquad \text{for branch 1} \\ \Psi_2(x) &= +\sqrt{u} \qquad \text{for branch 2} \\ \left|\frac{dx}{du}\right| &= \frac{1}{2\sqrt{u}} \qquad \forall\, x. \end{aligned}$$

For $0 \le u \le 9$ we must take into account both branches,

$$\begin{aligned} g(u) = P(U{=}\Phi(X)) &= \left[P(X = -\sqrt{U}) + P(X = +\sqrt{U})\right] \left|\frac{dx}{du}\right| \\ &= \left[\frac{1}{81}(-\sqrt{u})^2 + \frac{1}{81}(+\sqrt{u})^2\right] \cdot \frac{1}{2\sqrt{u}} = \frac{\sqrt{u}}{81}, \end{aligned}$$

while for $9 \le u \le 36$ there is only one branch,

$$\begin{aligned} g(u) = P(U{=}\Phi(X)) &= P(X = +\sqrt{U}) \cdot \left|\frac{dx}{du}\right| \\ &= \frac{1}{81}(+\sqrt{u})^2 \cdot \frac{1}{2\sqrt{u}} = \frac{\sqrt{u}}{2 \cdot 81}. \end{aligned}$$

4.2.2 Two or more variables

- The transformation from two variables (X, Y) to (U, V), all in continuous sampling spaces, proceeds by the same argument that, for some small rectangles $\Delta\mathcal{A}(x, y) = (x, x + \Delta x) \otimes (y, y + \Delta y)$ and $\Delta\mathcal{A}(u, v) = (u, u + \Delta u) \otimes (v, v + \Delta v)$, the interval probabilities must be the same

$$P(x \le X \le x + \Delta x, y \le Y \le y + \Delta y) = P(u \le U \le u + \Delta u, v \le V \le v + \Delta v) \tag{4.23}$$

or in terms of probability densities

$$p(x, y)\, \Delta x\, \Delta y = g(u, v)\, \Delta u\, \Delta v. \tag{4.24}$$

Following the same steps and arguments as in the one-variable case above, we obtain a transformation law involving the absolute value of the Jacobi determinant J,

$$\int_{\Delta\mathcal{A}(x,y)} dx\, dy\, p(x, y) = \int_{\Delta\mathcal{A}(u,v)} du\, dv\, g(u, v) = \int_{\Delta\mathcal{A}(x,y)} dx\, dy\, |J|\, g(u, v) \tag{4.25}$$

where in the second step we changed variables back from $(du\, dv)$ to $(dx\, dy)$ and where

$$J = \det \begin{bmatrix} \dfrac{\partial u}{\partial x} & \dfrac{\partial u}{\partial y} \\ \dfrac{\partial v}{\partial x} & \dfrac{\partial v}{\partial y} \end{bmatrix} = \frac{\partial(u, v)}{\partial(x, y)}. \tag{4.26}$$

Since the integral in (4.25) runs over x and y, we must substitute the forward transformations $u = \Phi_1(x, y)$ and $v = \Phi_2(x, y)$ in g to obtain

$$\int_{\Delta\mathcal{A}(x,y)} dx\, dy\, p(x, y) = \int_{\Delta\mathcal{A}(x,y)} dx\, dy\, |J|\, g[u{=}\Phi_1(x, y), v{=}\Phi_2((x, y)] \tag{4.27}$$

and so

$$p(x,y) = g[u{=}\Phi_1(x,y), v{=}\Phi_2(x,y)] \left|\frac{\partial(u,v)}{\partial(x,y)}\right| \tag{4.28}$$

The Jacobian determinants for the forward and inverse transformation are related simply by

$$\left|\frac{\partial(x,y)}{\partial(u,v)}\right| = \left|\frac{\partial(u,v)}{\partial(x,y)}\right|^{-1} \tag{4.29}$$

and the differential transformation law is therefore, in terms of the inverse transformations Ψ_1 and Ψ_2,

$$g(u,v) = p[x{=}\Psi_1(u,v), y{=}\Psi_2(u,v)] \cdot \left|\frac{\partial(x,y)}{\partial(u,v)}\right|. \tag{4.30}$$

- Again, care must be taken in the transformation of the outcome space. The process may involve several steps depending on whether the boundaries are simple and on whether the transformation is linear or nonlinear. We show by example of a linearly bounded $\mathcal{A}(X,Y)$ and linear transformations how the simplest case may work. Let

  $\mathcal{A}(X,Y)$ = triangle bounded by $\quad (X_A,Y_A) = (-1,0), \quad (X_B,Y_B) = (+1,0), \quad (X_C,Y_C) = (0,2)$

  $$U = \Phi_1(X,Y) = X + Y \qquad V = \Phi_2(X,Y) = Y$$

  $$X = \Psi_1(U,V) = U - V \qquad Y = \Psi_2(U,V) = V$$

  The three vertices transform to

  $$(X_A,Y_A) \to (U_A,V_A) = (-1,0) \qquad (X_B,Y_B) \to (U_B,V_B) = (1,0) \text{ and}$$
  $$(X_C,Y_C) \to (U_C,V_C) = (2,2).$$

  which is a skewed triangle in the (U,V) plane. Due to the linearity of the transformation, we know that the edges of $\mathcal{A}(U,V)$ are straight lines. To belabour the point, let us nevertheless calculate the transformation of the line AC:

  $$y = 2x + 2 \qquad\qquad v = y = 2(u - v) + 2 \qquad\qquad v = \tfrac{2}{3}u + \tfrac{2}{3}.$$

  which correctly describes the line between A' and C'.

- A final note on two-dimensional transformations for both discrete and continuous multivariate $\mathcal{A}$: With the possible exception of LI convolutions treated below, it is important always to do the full transformation $(X,Y) \to (U,V)$, even if only one of the final variables U or V is of interest. sampling spaces and probabilities of U alone or V alone must be determined as marginals of the two-dimensional probability and sampling space.

4.3 Convolutions and correlation functions

For our purposes, there are two important special cases of transformations followed by finding the marginal distribution: convolutions and autocorrelation functions. We consider both briefly.

Convolution: Suppose we have two variables X and Y but are interested only in their sum $U = X+Y$ and do not care about them separately. The general convolution, which is simply the marginal distribution $g(u)$, is determined by doing a full $(X,Y) \to (U,V)$ transformation and then integrating

out V. For the convolution, this can be accomplished by the skew transformation $U = X + Y$ and $V = Y$ with inverse $X = U - V$ and $Y = V$ which, since $|J| = 1$ yields, quite generally

$$g(u,v) = p(x{=}u{-}v, y{=}v)\,|J| = p(u{-}v, v) \tag{4.31}$$

$$g(u) = \int_{\mathcal{A}(V\,|\,u)} dv\, p(u{-}v, v). \tag{4.32}$$

In the special case where X and Y are logically independent i.e. $p(x,y) = p_1(x)\,p_2(y)$, this can be simplified to

$$g(u) \overset{\text{LI}}{=} \int_{\mathcal{A}(V\,|\,u)} dv\, p_1(u-v)\,p_2(v). \tag{4.33}$$

For discrete variables, the convolution is

$$g(u) \overset{\text{LI}}{=} \sum_{v\in\mathcal{A}(V\,|\,u)} p_1(u-v)\,p_2(v). \tag{4.34}$$

Once again, it is important to note that for finite sampling spaces $\mathcal{A}(U,V)$ does not in general equal the product $\mathcal{A}(U){\otimes}\mathcal{A}(V)$ even if $\mathcal{A}(X,Y) = \mathcal{A}(X){\otimes}\mathcal{A}(Y)$. For example, if X and Y have the same sampling space $\mathcal{A}(X) = \mathcal{A}(Y) = [A,B]$, then the above skew transformation from the square $\mathcal{A}(X,Y)$ with corners (A,A), (B,A), (A,B) and (B,B) results in a joint $\mathcal{A}(U,V)$ in the form of a rhombus with corners $(2A,A)$, $(A{+}B,A)$, $(A{+}B,B)$ and $(2B,B)$. Integrating out v must then take this rhombus form into account for every given value of u.

Autocorrelation functions play a crucial role in Markov Chain Monte Carlo methods. While in the case of a convolution, we are interested in the sum $X{+}Y$, in the case of general correlation functions we are interested in the difference $X{-}Y$. The formal mathematics is therefore quite similar. We define the skew transformation $U = X{-}Y$ and $V = Y$ so that

$$g(u,v) = p(x{=}u{+}v, y{=}v)\,|J| = p(u{+}v, v) \tag{4.35}$$

$$g(u) = \int_{\mathcal{A}(V\,|\,u)} dv\, p(u{+}v, v) \overset{\text{LI}}{=} \int_{\mathcal{A}(V\,|\,u)} dv\, p_1(u{+}v)\,p_2(v) \tag{4.36}$$

where the marginal distributions p_1 and p_2 are not in general the same functions. The case then they are equal, $p_1 = p_2$, corresponds to the case where we are *correlating* the variable V to a second variable $V{+}u$ of the same pdf, integrating over all V while the difference u is kept constant. This is then used to construct the *autocorrelation function*

$$C'(u) := \int_{\mathcal{A}(V\,|\,u)} dv\, p(v{+}u)\,p(v) \qquad \forall\, u. \tag{4.37}$$

Typically, however, the autocorrelation function is defined in analogy to the covariance $E(uv) - E(u)\,E(v)$ by subtracting the separate expectation values, so that the autocorrelation function for our purposes is

$$C(u) := \left[\int_{\mathcal{A}(V\,|\,u)} dv\, p(v{+}u)\,p(v)\right] - \left[\int_{\mathcal{A}(V\,|\,u)} dv\, p(v{+}u)\right]\left[\int_{\mathcal{A}(V\,|\,u)} dv'\, p(v')\right] \tag{4.38}$$

For discrete variables, the autocorrelation function is

$$C(u) := \left[\sum_{v\in\mathcal{A}(V\,|\,u)} p(v{+}u)\,p(v)\right] - \left[\sum_{v\in_{\mathcal{A}(V\,|\,u)}} p(v{+}u)\right]\left[\sum_{v'\in\mathcal{A}(V\,|\,u)} p(v')\right] \tag{4.39}$$

Chapter 5

Expectation values and generating functions

This chapter continues the exposition of general probability theory which applies to both deductive and inductive calculations. Expectation values start as simple characterisations of probabilities and acquire increasing importance in the form of generating functions and eventually in the central limit theorem.

5.1 Expectation values

- The simplest expectation value is that of x,

$$\mu_1 = E(x) = \int_{\mathcal{A}(X)} dx\, x\, p(x)$$

or, for discrete sampling spaces, $\mu_1 = \sum_{x\in\mathcal{A}(X)} x\, p(x)$. This forms part of the family of expectation values called *moments of order q*

$$\mu_q = E(x^q) = \int_{\mathcal{A}(X)} dx\, x^q\, p(x) \tag{5.1}$$

or an equivalent sum for discrete x. The variable q can theoretically take on any real value for which the integral or sum is finite; in practice, $q \in \mathbb{N}_0 = \{0, 1, 2, \ldots\}$.

- In general, the expectation value of any function $h(x)$ defined and bounded over $\mathcal{A}(X)$ is

$$E[h(x)\,|\,p(x)] = \begin{cases} \int_{\mathcal{A}(X)}\ dx\, h(x)\, p(x) \\ \sum_{x\in\mathcal{A}(X)}\ h(x)\, p(x), \end{cases} \tag{5.2}$$

where we have introduced a conditional notation in the definition of $E[\cdot]$ to emphasise its dependence on the particular probability on which the expectation value is taken. This also emphasises that expectation values depend on either prior knowledge of $p(x)$ and of course $\mathcal{A}(X)$, as encountered in deductive probability theory, or on the use of $p(x)$ in the context of an inferential hypothesis.

- Mathematically, the expectation value is a functional over $h(x)$ and $p(x)$ (i.e. resulting in a single number or value) but can also be seen as a function of any parameters or other variables which are not integrated out.

5.2 Expectation values under transformation

- Expectation values have simple properties under transformation. Again let $u = \Phi(x)$ with probability $g(u)$ and sampling space $\mathcal{A}(U)$. For discrete outcomes Eq. (4.5) tells us that $g(u) = \sum_{X\in\mathcal{A}(X)} P(X{=}x)\,\delta(u,\Phi(x))$ and so, on interchanging sums

$$E(u) = \sum_{u\in\mathcal{A}(U)} u\,g(u) = \sum_u u\left[\sum_x p(x)\,\delta(u,\Phi(x))\right] = \sum_x p(x)\left[\sum_u u\,\delta(u,\Phi(x))\right] \tag{5.3}$$

$$= \sum_x p(x)\,\Phi(x) = E[\Phi(x)] \tag{5.4}$$

or for continuous outcomes using Eq. (4.18)

$$E[u] = \int_{\mathcal{A}(U)} du\,u\,g(u) = \int_{\mathcal{A}(U,X)} du\,u\,dx\,\delta(u-\Phi(x))\,p(x) = \int_{\mathcal{A}(X)} \Phi(x)\,p(x)\,dx = E[\Phi(x)];$$

in other words, the expectation value of u with respect to $g(u)$ is the same as the expectation value of the function $\Phi(x)$ with respect to $p(x)$. In short, expectation values can be taken over $g(u)$ or over $p(x)$, as long as we remember to substitute $\Phi(x)$ for u when doing the expectation value in $p(x)$,

$$\boxed{E(u) = E(\Phi(x))} \tag{5.5}$$

or in more detailed notation

$$E[u\,|\,g(u)] = E[\Phi(x)\,|\,p(x)] \qquad \text{or} \qquad E_u[u] = E_x[\Phi(x)] \tag{5.6}$$

but mostly the context will tell us which probability is involved, so we just use E.

- The transformation can of course be extended to expectations over any function $h(u)$. The most general result is therefore

$$\boxed{E[h(u)\,|\,g(u)] = E[h(\Phi(x))\,|\,p(x)]} \tag{5.7}$$

and choosing $h = \Phi^{-1}$ we derive immediately the inverse identity for $E(x)$ in terms of $E(\Phi^{-1}(u))$.

- When $\Phi(x)$ is nonbijective, $\sum_x$ must again be taken separately over every branch.Here is a quick example for discrete variables. Let $U = X^2 = \Phi(X)$,

$$\mathcal{A}(X) = \{-2,-1,0,1,2\} \qquad\qquad \mathcal{A}(U) = \{0,1,4\}$$
$$p(x) \in \{\tfrac{1}{8},\tfrac{1}{8},\tfrac{1}{8},\tfrac{3}{8},\tfrac{2}{8}\} \qquad\qquad P(U{=}0) = P(X{=}0) = \tfrac{1}{8}$$
$$P(U{=}1) = P(X{=}-1) + P(X{=}+1) = \tfrac{4}{8}$$
$$P(U{=}4) = P(X{=}-2) + P(X{=}+2) = \tfrac{3}{8}$$

and the corresponding expectation values are

$$\begin{aligned} E(u) &= \sum_{u=0,1,4} u\,g(u) = 0\cdot\tfrac{1}{8} + 1\cdot\tfrac{4}{8} + 4\cdot\tfrac{3}{8} \\ &= 0^2\cdot p(0) + 1^2\,[p(-1)+p(1)] + 2^2\cdot[p(-2)+p(2)] \\ &= 2^2\cdot p(-2) + 1^2\cdot p(-1) + 0^2\cdot p(0) + 1^2\cdot p(1) + 2^2\cdot p(2) = \sum_{x=-2}^{+2} \Phi(x)\,p(x) \\ &= E(x^2) \end{aligned}$$

5.3 Lemmas on expectation values

We briefly consider the behaviour of expectation values for the simplest transformations. We quote only the continuous versions; the equivalent discrete ones are easily written down.

- For constant c and any two variables X and Y, and remembering that the meaning of E depends on its argument,

$$E(c) = \int_{\mathcal{A}(X)} dx\, p(x)\, c = c \qquad \text{since} \qquad \int_{\mathcal{A}(X)} dx\, p(x) = 1 \tag{5.8}$$

$$E(cx) = c\, E(x) \tag{5.9}$$

$$E(x + c) = E(x) + c \tag{5.10}$$

$$E(x + y) = E(x) + E(y) \tag{5.11}$$

$$E(xy) \overset{\text{LI}}{=} E_x(x)\, E_y(y) \tag{5.12}$$

  A special case of the first identity is $E(E(x)) = E(x)$. The first three identities are easily proven. The fourth is more subtle, because it requires a PDF of two variables and several steps:

$$\begin{aligned} E(x+y) &= \int dx\, dy\, p(x,y)\, (x+y) = \int dx\, dy\, p(x,y)\, x + \int dx\, dy\, p(x,y)\, y \\ &= \int dx\, p_1(x)\, x \;+\; \int dy\, p_2(y)\, y \;=\; E_x(x) + E_y(y) \end{aligned}$$

  where the latter expectation values are taken with respect to the marginal probabilities $p_1(x)$ and $p_2(y)$. Note that this theorem is true with or without independence. By contrast, the factorisation theorem $E(xy) = E_x(x)\, E_y(y)$ is true only if X and Y are independent,

$$\begin{aligned} E(xy) &= \int_{\mathcal{A}(X,Y))} dx\, dy\, p(x,y)\, xy \\ &\overset{\text{LI}}{=} \int_{\mathcal{A}(X)\otimes\mathcal{A}(Y)} dx\, dy\, p_1(x)\, p_2(y)\, xy \\ &= \int_{\mathcal{A}(X)} dx\, p_1(x)\, x \cdot \int_{\mathcal{A}(Y)} dy\, p_2(y)\, y = E_x(x)\, E_y(y) \end{aligned}$$

  Note that the theorem holds only if the sampling space factorises, as it must for independent variables. Also note that, if any joint function $h(x, y)$ does not factorise into factors $h_1(x)$ and $h_2(y)$, then $E(h(x, y))$ also does not factorise, even if X and Y are independent.

5.4 Theoretical variance and standard deviation for X

- Having already dealt with most relevant issues, we can be brief. For the purposes of later reference, we first introduce the *cumulant* of first order. It is identical to the first moment: $\kappa_1 = \mu_1$.

- The theoretical variance κ_2 is defined as the expectation of the squared difference between the variable and its expectation; written in four equivalent ways, it is

$$\begin{aligned} \kappa_2 &= \text{var}(x) \\ &= E[(x - E(x))^2] \\ &= E[(x - \mu_1)^2] \\ &= \int dx\, (x - \mu_1)^2\, p(x) \end{aligned} \tag{5.13}$$

with a corresponding definition for discrete X. It can also be written as

$$\kappa_2 = \mu_2 - \mu_1^2 \;=\; E(x^2) - E(x)^2, \tag{5.14}$$

which follows from application of the above identities,

$$\begin{aligned} E[(x - E(x))^2] &= E[x^2 - 2E(x)\,x + E(x)^2] \\ &= E[x^2] + E[-2E(x)\,x] + E[E(x)^2] \\ &= E(x^2) - 2E(x)E(x) + E(x)^2 \end{aligned}$$

- **Interpretation**
  - For probability density functions $p(x)$ which are reasonably behaved, i.e. are unimodal[1] and are not too asymmetric, $\kappa_1{=}\mu_1$ is a measure of the *location* of the peak. It may, but does not have to, coincide with the exact position of the peak, but will in general be close to it.
  - The standard deviation $\sigma_x = \sqrt{\kappa_2}$ is a generic measure of the width of the pdf, which is a *scale*. A particular fraction of the probability or area under $p(x)$ is normally within the interval $\mu_1 - \sigma_x \leq x \leq \mu_1 + \sigma_x$.
  - As already stated, μ_1, κ_2 and σ_x depend on prior knowledge of $p(x)$. They are theoretical values and not data. The only connection with data is that the sample means and sample moments treated in Chapter 6 will tend asymptotically to these theoretical values. One can therefore make a pseudo-frequentist argument that μ_1 would be the result for sample means when the number of measurements tend to infinity, but this is not a necessity.

5.5 Theoretical moments, variances and covariance for (X, Y)

- We again consider continuous variables and probability densities only; the discrete equivalents are readily written down. Given a bivariate pdf $p(x, y)$ with marginals $p_1(x)$ and $p_2(y)$, we can define moments with two exponents q_1, q_2 as

$$\mu_{q_1,q_2} = E(x^{q_1} y^{q_2}) = \int dx\, dy\, x^{q_1} y^{q_2}\, p(x, y). \tag{5.15}$$

- While higher-order moments for q_1 and q_2 are certainly of importance later, we are currently concerned only with the lowest-order ones of $(q_1, q_2) = (0,1), (1,0), (0,2), (2,0), (1,1)$. Of the many notations that exist, we here consider only two:
  - The subscripts are replaced by the variables, for example $\mu_{0,2} \to \mu_{y,y}$ and $\mu_{1,1} \to \mu_{x,y}$. This notation is useful for extension to many variables $\boldsymbol{X} = (X_1, X_2, \ldots)$ and easily maps onto the indices of so-called *covariance matrices*.
  - For the purposes of transformations in general and convolutions and autocorrelations in particular, a notation which places the transformation or variable of interest in brackets as an argument of the moment is more appropriate. We give a few examples below of both notations and how they are related.
  - In addition, the traditional notations var and cov are also included because they are often encountered in the literature.

[1] A probability is unimodal if it has only a single maximum (peak) and the maximum is not located on the boundary of the sampling space.

- Here is a summary of the most common forms in their various notations

$$\kappa_x = \kappa_1(x) = \mu_1(x) = \int dx\, dy\, p(x,y)\, x = \int dx\, p_1(x)\, x \tag{5.16}$$

$$\kappa_y = \kappa_1(y) = \mu_1(y) = \int dx\, dy\, p(x,y)\, y = \int dy\, p_2(y)\, y \tag{5.17}$$

$$\mu_{x,x} = \mu_2(x) = \int dx\, dy\, p(x,y)\, x^2 = \int dx\, p_1(x)\, x^2 \tag{5.18}$$

$$\kappa_{x,x} = \kappa_2(x) = \mathrm{var}(x) = \Big[\int dx\, p_1(x)\, x^2\Big] - \Big[\int dx\, p_1(x)\, x\Big]^2 \tag{5.19}$$

$$\kappa_{y,y} = \kappa_2(y) = \mathrm{var}(y) = \Big[\int dy\, p_2(y)\, y^2\Big] - \Big[\int dy\, p_2(y)\, y\Big]^2 \tag{5.20}$$

$$\kappa_{x,y} = \mathrm{cov}(x,y) = \Big[\int dx\, dy\, p(x,y)\, xy\Big] - \Big[\int dx p_1(x)\Big]\Big[\int dy\, p_2(y)\, y\Big]. \tag{5.21}$$

We write the square roots of the second-order cumulants as

$$\sigma_x = \sqrt{\kappa_{x,x}} \qquad \sigma_y = \sqrt{\kappa_{y,y}} \qquad \sigma_{x,y} = \sqrt{\kappa_{x,y}} \tag{5.22}$$

The *standard deviations* σ_x represents a typical scale associated with the width of a unimodal probability $p(x)$; note that it can be understood both in terms of the marginal $p_1(x)$ of a higher-order pdf, or as the width of a single-variable probability density. The *variance* $\kappa_{x,x}$ is of course the square of the standard deviation. All these properties hold correspondingly for σ_y and $\kappa_{y,y}$.

- The *covariance* $\kappa_{x,y}$ is the generalisation of the variance,

$$\sigma^2_{x,y} = \kappa_{x,y} = \mathrm{cov}(x,y) = E[xy] - E[x]\,E[y] = E\big[(x - E(x))\,(y - E(y))\big] \tag{5.23}$$

It is zero when x and y are statistically independent since as shown above $E[xy] \overset{\text{LI}}{=} E[x]\,E[y]$.

- It is not hard to show that

$$\kappa_2(cx) = \mathrm{var}(cx) = c^2\kappa_2(x) \tag{5.24}$$

$$\kappa_2(x+y) = \kappa_2(x) + \kappa_2(y) + 2\,\mathrm{cov}(x,y) \tag{5.25}$$

$$\kappa_2(x-y) = \kappa_2(x) + \kappa_2(y) - 2\,\mathrm{cov}(x,y) \tag{5.26}$$

$$\mathrm{cov}(x,y) \le \sqrt{\kappa_2(x)\,\kappa_2(y)} \tag{5.27}$$

The latter is known as the Cauchy-Schwarz inequality.

- The **Pearson correlation coefficient** is defined as a "normalised" covariance

$$\rho(x,y) = \frac{\mathrm{cov}(x,y)}{\sqrt{\mathrm{var}(x)\ \mathrm{var}(y)}} = \frac{\sigma^2_{x,y}}{\sigma_x\,\sigma_y}. \tag{5.28}$$

It can be shown that the Pearson coefficient is just the covariance for the corresponding two standardised variables x^* and y^* considered below. From the fact that $\mathrm{var}(x^* + y^*)$ and $\mathrm{var}(x^* - y^*)$ cannot be negative, it can also be shown that

$$-1 \le \rho \le +1. \tag{5.29}$$

5.6 Standardised variables

- Given a pdf $p(x)$, we can usually calculate its μ_1 and standard deviation $\sqrt{\kappa_2}$. We then transform to the *standardised variable* $u = x^* = \Phi(x)$ with the transformation

$$x^* = \Phi(x) = \frac{x - \mu_1}{\sqrt{\kappa_2}} \tag{5.30}$$

- Using the above and the inverse transformation $x = \Phi^{-1}(x^*) = \sqrt{\kappa_2}\, x^* + \mu_1$, the old and new pdfs are related by

$$p(x) = g(u{=}\Phi(x)) \left|\frac{dx^*}{dx}\right| = \frac{1}{\sqrt{\kappa_2}}\, g\left(x^* = \frac{x-\mu_1}{\sqrt{\kappa_2}}\right) \tag{5.31}$$

$$g(x^*) = p(\Phi^{-1}(x^*)) \left|\frac{dx}{dx^*}\right| = \sqrt{\kappa_2}\; p(x = \sqrt{\kappa_2}x^*{+}\mu_1) \tag{5.32}$$

- The new $g(x^*)$ has a "standard form" in that the location and variance are 0 and 1. Using the above theorems and $\kappa_2 = \kappa_{x,x} = E[(x-\mu_1)^2]$, we find, with the help of (5.14),

$$E(x^*) = \int dx^*\, g(x^*)\, x^* = \int dx\, p(x) \left(\frac{x-\mu_1}{\sqrt{\kappa_2}}\right) = \frac{1}{\sqrt{\kappa_2}}\,[E(x) - \mu_1] = 0 \tag{5.33}$$

$$\mathrm{var}(x^*) = \kappa_2(x^*) = E\left[\left(\frac{x-\mu_1}{\sqrt{\kappa_2}}\right)^2\right] - E(x^*)^2 = \frac{1}{\sqrt{\kappa_2}^2}\, E\left[(x-\mu_1)^2\right] - 0 = \frac{\kappa_2}{\kappa_2} = 1 \tag{5.34}$$

- The transformation $(x-\mu_1)/\sqrt{\kappa_2}$ effects a *shift* of the pdf to the left by an amount μ_1, followed by a *squeeze* or *rescaling* of its width by a factor $\sqrt{\kappa_2}$.

- Standardisation is possible and useful for all continuous-x distributions which have a finite first and second moment. It is not limited to the Gaussian.

- By its very definition, x^* is always dimensionless.

5.7 The Gaussian pdf

- The above definitions are especially simple for the case of the gaussian distribution, which happens to have two parameters μ and σ such that

$$p(x\,|\,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \qquad \mathcal{A}(x) = \mathbb{R} \tag{5.35}$$

  for which

$$\kappa_1 = \mu_1 = E(x) = \mu \tag{5.36}$$

$$\kappa_2 = E(x^2) - E(x)^2 = \sigma^2. \tag{5.37}$$

  This close correspondence between the generally defined (κ_1, κ_2) on the one hand and the gaussian parameters (μ, σ) on the other hand leads to much confusion, because (κ_1, κ_2) exist for most other nongaussian pdfs also. For the gaussian, the parameter μ, the expectation value μ_1 and the maximum (called *mode*) are all the same. This is not generally true; μ_1 and κ_2 exist for many other distributions with other parameters and are not necessarily equal to any of them.

- The standardised form of the gaussian is

$$p(x^*) = \frac{e^{-(x^*)^2/2}}{\sqrt{2\pi}}. \tag{5.38}$$

- The *error function*, defined as

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z du\, e^{-u^2} \tag{5.39}$$

  is a partial integral of the standardised Gaussian.

- The following interval probabilities play an important role in interpretations using the gaussian:

$$\int_{\mu-\sigma}^{\mu-\sigma} dx\, p(x \,|\, \mu, \sigma) = \int_{-1}^{+1} dx^*\, p(x^*) = \mathrm{erf}(1/\sqrt{2}) = 0.6826\ldots \tag{5.40}$$

$$\int_{\mu-2\sigma}^{\mu-2\sigma} dx\, p(x \,|\, \mu, \sigma) = \int_{-2}^{+2} dx^*\, p(x^*) = \mathrm{erf}(2/\sqrt{2}) = 0.9545\ldots, \tag{5.41}$$

  i.e. they tell us how much of the total probability 1 is contained within one and within two standard deviations from μ_1.

- The error function is closely related to the distribution function for the standardised gaussian; using $u = x^*/\sqrt{2}$, we find for $z > 0$

$$F(z) = P(X^* \le z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} dx^*\, e^{-(x^*)^2/2} \tag{5.42}$$

$$= \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_0^z dx^*\, e^{-(x^*)^2/2} = \frac{1}{2} + \frac{1}{2}\frac{2}{\sqrt{2\pi}} \int_0^{z/\sqrt{2}} \sqrt{2}\, du\, e^{-u^2} \tag{5.43}$$

$$= \frac{1}{2}\left[1 + \mathrm{erf}(z/\sqrt{2})\right] \tag{5.44}$$

  a similar expression can be found for $z < 0$.

5.8 Moments and the moment generating function

We have in Eq. (5.1) already introduced the moments $\mu_q, q = 1, 2, 3, \ldots$ for a given probability density $p(x)$. While these can be calculated by doing the appropriate integrals, they are also derivatives of the so-called *moment generating function* (mgf), which we now discuss. Apart from its technical utility in calculating moments, the moment generating function is an essential tool both in convolution calculations and in statistical physics. On an even more fundamental level, the mgf and other generating functions open up a way to embed a theory into a higher-dimensional space, within which constraints are accommodated more easily.

The mgf is defined in terms of a new "dual variable" h as

$$M(h \,|\, p(x)) = E(e^{hx}) = \int_{\mathcal{A}(X)} dx\, p(x)\, e^{hx} \tag{5.45}$$

Notation: Most books will use the symbol t instead of the h we have used here, but because t denotes a time variable in stochastic processes and in general time-dependent contexts, we prefer to allocate a different symbol to the dual variable. As usual, we utilise the vertical line to indicate which quantities are known and fixed; in this case the probability density $p(x)$.

If the integral in (5.45) can be solved in closed form, then derivatives of the solution immediately give us the moments of $p(x)$ because

$$\frac{dM}{dh} = E\left(xe^{hx}\right) = \int dx\, p(x)\, x\, e^{hx}$$

so that the first moment is the derivative taken at h=0,

$$\left.\frac{dM}{dh}\right|_{h=0} = \int dx\, p(x)\, x\, e^{hx}\Big|_{h=0} \;=\; \mu_1 \tag{5.46}$$

and so the q-th derivative at zero equals the q-th moment,

$$\mu_q = \left.\frac{d^q M}{dh^q}\right|_{h=0} = E(x^q) = \int dx\, p(x)\, x^q. \tag{5.47}$$

Given the Taylor expansion around $h = 0$ (we suppress the $p(x)$ in M to simplify the notation)

$$M(h) = M(0) + h\left.\frac{dM}{dh}\right|_{h=0} + \frac{h^2}{2!}\left.\frac{d^2M}{dh^2}\right|_{h=0} + \dots \tag{5.48}$$

and $M(0) = 1$, we can write the generating function as the series

$$M(h) \;=\; 1 + h\,\mu_1 + \frac{h^2}{2!}\,\mu_2 + \frac{h^3}{3!}\,\mu_3 + \dots \;=\; \sum_{r=0}^{\infty} \frac{h^q}{r!}\,\mu_q \tag{5.49}$$

For the case of **two variables**, the m.g.f. is a function of two parameters h_1 and h_2,

$$M(h_1, h_2 \,|\, p(x_1, x_2)) = E\left(e^{h_1 x_1 + h_2 x_2}\right) = \int_{\mathcal{A}(X_1, X_2)} dx_1\, dx_2\, p(x_1, x_2)\, e^{h_1 x_1 + h_2 x_2}. \tag{5.50}$$

Moments for X_1, for X_2 separately as well as mixed moments which depend on both X_1 and X_2 can also be found directly from $M(h_1, h_2 \,|\, p(x_1, x_2))$. Such moments must, of course, be written as μ_{q_1,q_2} in terms of two orders q_1 and q_2 which indicate the respective exponents of x_1 and x_2, or in some tensor notation such as μ_{x_1,x_2} similar to the notation used in Section 5.4.

If X_1 and X_2 are independent, (i.e. $p(x_1, x_2) \overset{\text{LI}}{=} p_1(x_1)\, p_2(x_2)$ and $\mathcal{A}(X_1, X_2) \overset{\text{LI}}{=} \mathcal{A}(X_1) \otimes \mathcal{A}(X_2)$), the m.g.f. factorises

$$\begin{aligned} M(h_1, h_2 \,|\, p(x_1, x_2)) &\overset{\text{LI}}{=} \int_{\mathcal{A}(X_1)} dx_1 \int_{\mathcal{A}(X_2)} dx_2\, p_1(x_1)\, p_2(x_2)\, e^{h_1 x_1 + h_2 x_2} \\ &\overset{\text{LI}}{=} M(h_1 \,|\, p_1(x_1)) \;\cdot\; M(h_2 \,|\, p_2(x_2)). \end{aligned} \tag{5.51}$$

Mgf for the sum: The special case $h_1 = h_2 = h$ is the m.g.f. for the sum $(X_1 + X_2)$,

$$M(h \,|\, p(x_1, x_2)) = E(e^{h(x_1 + x_2)}) = \int_{\mathcal{A}(X_1, X_2)} dx_1\, dx_2\, p(x_1, x_2)\, e^{h(x_1 + x_2)} \tag{5.52}$$

since derivatives with respect to h at h=0 yield

$$\mu_q(x_1{+}x_2) = E((x_1{+}x_2)^q). \tag{5.53}$$

where we have added the argument $(x_1{+}x_2)$ to μ_q to emphasise that this μ_q is the moment of a sum. (Actually, we should write something like $\mu(q, x_1 + x_2)$, but the notation becomes too cumbersome.) The factorisation under independence also holds for the sum-m.g.f. with $h_1 = h_2 = h$,

$$M(h \,|\, p(x_1, x_2)) \overset{\text{LI}}{=} \int_{\mathcal{A}(X_1)} dx_1 \int_{\mathcal{A}(X_2)} dx_2\, p_1(x_1)\, p_2(x_2)\, e^{h x_1 + h x_2} \overset{\text{LI}}{=} M(h \,|\, p_1(x_1))\; M(h \,|\, p_2(x_2)). \tag{5.54}$$

Note that in the latter case the *same* h now appears in both m.g.f.'s on the right hand side.

From this we can generalise to the sum of N independent variables $x_1, x_2, \ldots, x_N$. Under the strong condition of N-fold independence

$$p(\boldsymbol{x}) = p(x_1, x_2, \ldots, x_N) \stackrel{\text{LI}}{=} \prod_{n=1}^{N} p_n(x_n) \tag{5.55}$$

the generating function $M(S \,|\, p(\boldsymbol{x}))$ of the sum $S = \sum_{n=1}^{N} x_n$ is the product of the individual generating functions $M(h \,|\, p(x_n))$,

$$M(S \,|\, p(\boldsymbol{x})) = \prod_{n=1}^{N} M(h \,|\, p_n(x_n)) \tag{5.56}$$

and if all the individual pdfs have the same form, $p_n(x_n) = p(x_n)$, it becomes the Nth power of the individual mgf,

$$M(S \,|\, p(\boldsymbol{x})) = [M(h \,|\, p(x_1)]^N \tag{5.57}$$

Thus, the *moment generating function of a convolution of N independent variables equals the N-th power of the mgf of a single variable.* This theorem forms the mathematical basis of the statements in statistical physics that the total partition function of N systems equals the N-th power of the partition function of one system.

We note in passing that the moment generating function is closely related to the so-called *characteristic function* (cgf), which is defined as

$$\Phi(z) = E[e^{izx}] \tag{5.58}$$

and where Φ refers not to a transformation but to the cgf.

5.9 The cumulant generating function

The m.g.f is closely related to the so-called **cumulant generating function** (c.g.f.) as its logarithm,

$$K(h \,|\, p(x)) \equiv \ln M(h \,|\, p(x)) = \ln E(e^{hx}) \tag{5.59}$$

and it has a Taylor expansion in terms of **cumulants** κ_q,

$$K(h \,|\, p(x)) = \kappa_1 h + \kappa_2 \frac{h^2}{2!} + \kappa_3 \frac{h^3}{3!} + \ldots \quad = \sum_{q=1}^{\infty} \kappa_q \frac{h^q}{q!} \,. \tag{5.60}$$

As in the previous section, we shall sometimes write $\kappa_q(x)$ although the cumulant is not a direct function of x but composed of integrals over $p(x)$. Note that, in contrast to the expansion of the mgf, there is no q=0 term in the sum. Because the κ_q are coefficients of the Taylor expansion, we have

$$\kappa_q = \frac{d^q}{dh^q} K(h \,|\, p(x)) \bigg|_{h=0} \,. \tag{5.61}$$

We will often omit the p in this notation and write $K(h \,|\, x)$ for the above and similar cases.

5.10 Properties of cumulants and the cgf

Cumulants and the corresponding cumulant generating functions have the following important properties:

(a) Relationship to moments:

$$\begin{aligned} \kappa_1 &= \mu_1 = E[x] \\ \kappa_2 &= \mu_2 - \mu_1^2 = \text{var}(x) = E\left[(x - E[x])^2\right] \\ \kappa_3 &= \mu_3 - 3\mu_2\mu_1 + 2\mu_1^3 \qquad \text{and so on} \end{aligned}$$

(b) Additivity under LI of cumulant generating function

$$K(h_1, h_2 \,|\, p(x_1, x_2)) \overset{\text{LI}}{=} K(h_1 \,|\, p_1(x_1)) + K(h_2 \,|\, p_2(x_2)), \tag{5.62}$$

(c) Additivity under LI of sum of variables,

$$K(h \,|\, g(x_1 + x_2)) \overset{\text{LI}}{=} K(h \,|\, p_1(x_1)) + K(h \,|\, p_2(x_2)) \tag{5.63}$$

$$\kappa_q(x_1 + x_2) \overset{\text{LI}}{=} \kappa_q(x_1) + \kappa_q(x_2) \tag{5.64}$$

$$\kappa_q(\textstyle\sum_i x_i) \overset{\text{LI}}{=} \displaystyle\sum_i \kappa_q(x_i), \tag{5.65}$$

The property of additivity under LI gives cumulants them their name, because they "accumulate" the statistical properties of independent variables. Note that moments are not additive at all: $\mu_q(x_1 + x_2) \neq \mu_q(x_1) + \mu_q(x_2)$.

(d) Tensor properties under affine transformation, $U = cX + d$ with c, d constants,

$$K(h \,|\, g(u)) = hd + K(ch \,|\, p(x)) \tag{5.66}$$

$$\kappa_1(U) = d + c\,\kappa_1(x) \tag{5.67}$$

$$\kappa_q(U) = c^q \kappa_q(x) \quad \forall\, r = 2, 3, \ldots \tag{5.68}$$

(e) For the gaussian, cumulants of higher orders are zero; see below.

(f) It can also be shown that the covariance is equal to the bivariate cumulant $\text{cov}(x, y) = \kappa_{1,1} = E(xy) - E(x)\,E(y)$ and equivalent quantities of higher orders can also be defined.

Chapter 6

Data and statistics

6.1 Data

- Statistics is by now highly mathematical, but it started humbly from the need to quantify and explain data. We use the term *statistics* as pertaining to the present data-driven context, while *probability theory* pertains to the theoretical framework. Of course the two must eventually be unified into one system.
- What is data? From the point of view developed so far, a single answer is a "datum"; it is an actual answer obtained to a question asked. Generically, we called the question the variable X, and a datum corresponds to x. We proceed from uncertainty via a datum to certainty.
- At this level, there is as yet no list of possible answers and no probability: we simply have X and x. Often the matter ends there: question asked and answered. Only in cases where one would like to predict some future answer to this question would the next level of model-construction in the form of $\mathcal{A}$ and P be needed.
- Most of the time, there is more than one answer, i.e. we have an entire set of facts at our disposal which we call our dataset or sample set. Often, but not always, these result from repetition of the same question or experiment. Here are some typical examples:
  - **Die:** Tossing the die 10 times, we may obtain a sample set such as $\{1, 4, 3, 3, 6, 2, 3, 1, 5, 2\}$.
  - **Your set of measurements:** For the purposes of statistics, a set of measurements is considered a sample set.
  - The set of **Monte Carlo results** of a computer simulation, normally after a repetition of some algorithm (e.g. end points of random walks, number of random numbers in a given interval etc.) To distinguish them from real data, we sometimes call simulation results **simdata**.
- A sample set is a finite number of individual samples $\mathcal{D} = \{x_1, x_2, \ldots, x_N\} = \{x_i\}_{i=1}^N$, where N is the number of elements in the sample which were obtained. Traditionally, the entire set would have been called "a sample"; modern usage refers to each single x_i as a sample.
- There are of course many issues of knowing which question to ask, which kind of data to collect, how to collect it, whether that collection is accurate and so on. **Data acquisition** and everything related to it is a big issue. While this can all be addressed within the Bayesian framework, we assume in this course that such issues have been resolved.
- In physics, we normally deal with numerical rather than qualitative data. This comes in two basic forms. In case of *raw data* $\mathcal{D} = \{x_i\}_{i=1}^N$, the individual x_i could be real numbers, rational numbers, integers etc.

- Raw data can be viewed and visualised in **scatter plots**, in which each x_i is plotted as a point (or, when there is only one variable, as a little vertical line) on the real line.

6.2 Binning and data counts

- In the case of raw data $\{x_i\}$, we must further distinguish between *index-ordered* and *-unordered* cases. Ordering matters if the subscript i of the data x_i carries information of significance such as time, in which case we must decide whether such acquisition time information is important or can be discarded.

- If the subscript i is used for indexing purposes only but has no further significance, it is normally a good idea to convert the raw data $D = \{x_i\}_{i=1}^{N}$ into a set of counts $\mathcal{D}' = \{n_b\}_{b=1}^{B}$ by means of a so-called *binning* procedure as set out in more detail below. This can help to strongly compress the amount of data which has to be stored.

- Before raw data can be converted to counts, two tasks must be accomplished: firstly, we must decide what values x_i could be possible, and secondly we must subdivide the set of all those possible values into bins.

  It seems easy enough to find the real number interval $\mathcal{A}$ within which all data $\mathcal{D}$ will fall: simply find the minimum and maximum of the set. While this works for quick-and-dirty processing, there are some issues to think about.

  - The first is simply to augment the interval by extending those minimum and maximum readings into human-readable numbers; given $(x_{\min}, x_{\max}) = (0.15332423, 3.51218760)$ we would be tempted to choose an interval $(0.00, 3.60)$.
  - It often happens, however, that a particular data sample is not unique. Choosing different intervals for each is possible but unhelpful for comparison. Choosing only a single interval carries the danger that some future datum may fall outside our chosen range. On the other hand choosing a huge interval just to accommodate all such possibilities may result in lots of empty space on our graphs.
  - In practice, the best solution is to choose a reasonable interval for $\mathcal{A}$ but to define so-called *underflow* and *overflow* bins. Any future x_i which falls outside the boundaries of $\mathcal{A}$ is then counted in these "emergency" bins, and while the analysis will continue, good software will inform the user that the chosen $\mathcal{A}$ was breached.

- Once the overall sampling space has been determined, $\mathcal{A}$, we can proceed to partition it. The set theory of partitions has already been discussed in Section 3.1.2.

- For cases with **discrete elementary outcomes** and relatively few possible outcomes, it is always possible to keep each of these as a "bin". For data of die tosses, we would normally "partition" the sample space into $\mathcal{A}_1 = \{1\}, \mathcal{A}_2 = \{2\}, \cdots, \mathcal{A}_6 = \{6\}$.

  In other cases, combining several discrete elementary outcomes into a single bin may make more sense. In general, there is no requirement that the $\mathcal{A}_b$'s must have the same size or be topologically simple. Nonetheless, we often do try to make bin sizes equal and we often do order the bins, e.g. for the die we could have a partition with equal bin widths of 2 and with ordered outcomes $\mathcal{A}_1 = \{1, 2\}, \mathcal{A}_2 = \{3, 4\}, \mathcal{A}_3 = \{5, 6\}$.

- For **real elementary outcomes**, the corresponding sampling space cannot of course list each possible value of x_i; rather, we define some continuous line interval or part of the real line. Again, the partition of $\mathcal{A}$ into a set of bins can be chosen by the user, and the different bin sizes $\mathcal{A}_b$ need not be equal. There are cases where bin sizes changing exponentially are useful.

The maths for the case where bins do have equal size can be formulated generally. If e.g. $\mathcal{A}(X) = [A, A{+}L]$ is the real line interval between A and $A{+}L$, we normally define a set of B bins each with a width $|\mathcal{A}_b| = \eta = L/B$. The bin boundaries are A plus $0, \eta, 2\eta, \ldots, b\eta, \ldots, (B{-}1)\eta$. The partition is then the collection of half-open subsets (bins)

$$\mathcal{A}_b = \{x \,|\, A + (b-1)\eta \le x < A + b\eta\}, \qquad b = 1, 2, \ldots, B, \tag{6.1}$$

and the bin mid-points are given by $x_b = A + (b - \frac{1}{2})\eta$.

- The data count n_b in bin b is defined as the number of data points x_i which fall into a particular bin $\mathcal{A}_b$. Algorithmically, this is effected by running over all i and incrementing by 1 the count in that bin b for which $x_i \in \mathcal{A}_b$,

$$n_b = \sum_{i=1}^{N} \delta(x_i \in \mathcal{A}_b) \tag{6.2}$$

where the "logical delta function" $\delta(x_i \in \mathcal{A}_b)$ is a generalisation of the Kronecker delta notation which equals 1 whenever the logical condition in the argument is true.

- For a given sample of discrete outcomes, the *relative frequency* r_b for a given partition element $\mathcal{A}_b$ is defined as

$$r_b \equiv \frac{n_b}{N} \tag{6.3}$$

$$\sum_{b=1}^{B} r_b = 1 \tag{6.4}$$

- By contrast, for continuous outcomes we define the *relative frequency density*

$$r_b \equiv \frac{1}{\eta}\,\frac{n_b}{N} \tag{6.5}$$

The area enclosed by each bin is $\eta\, r_b = (n_b/N)$. The total area under the histogram is therefore

$$\sum_{b=1}^{B} \eta\, r_b = 1. \tag{6.6}$$

Since the bin width η may not be small, the relative frequency r_b is plotted at the centre of the bin, i.e. at the bin midpoint x_b.

- The set of counts, relative frequencies or relative frequency densities are visualised with a *histogram* which generally plots the set of r_b or n_b against the corresponding x_b, either as charts of unit-width bars for discrete x or as bars of width η for continuous x. Histograms therefore constitute a basic tool of data analysis.

6.3 Sample statistics from raw data

- It is often useful to describe a given sample set in a few numbers; each such descriptive number is called a *sample statistic*. One example of a sample statistic is the *sample mean*

$$m_1 = \langle x \rangle \equiv \frac{1}{N}\,(x_1 + \ldots + x_N) = \frac{1}{N}\sum_{i=1}^{N} x_i\,. \tag{6.7}$$

This is a special case of the *sample moment*

$$m_q = \langle x^q \rangle \equiv \frac{1}{N}\left(x_1^q + \ldots + x_N^q\right) = \frac{1}{N}\sum_{i=1}^{N} x_i^q \qquad r = 1, 2, \ldots \tag{6.8}$$

The *sample variance* is just as important; it is defined as

$$k_2(x) \equiv \langle x^2 \rangle - \langle x \rangle^2 = \left(\frac{1}{N}\sum_i x_i^2\right) - \left(\frac{1}{N}\sum_i x_i\right)\left(\frac{1}{N}\sum_j x_j\right) \tag{6.9}$$

$$= m_2 - m_1^2 \tag{6.10}$$

k_2 is also called the "second sample cumulant". Sample cumulants of higher orders can be found in a similar way directly from the data. Generally, the sample average of any function h of the raw data x_i is

$$\langle h(x) \rangle = \frac{1}{N}\sum_i h(x_i). \tag{6.11}$$

6.4 Sample statistics from relative frequencies

- Since the raw data can be re-ordered, the raw data can be grouped according to elementary outcome. On this basis, a second method of calculating sample statistics based on the relative frequency emerges as illustrated by the simple example of a die.

-
  - Random variable X = throw die once
  - sampling space $\mathcal{A} = \{1, 2, 3, 4, 5, 6\}$
  - Partition: normally we use the set of bins $\mathcal{A}_b = \{b\}, b = 1, 2, \ldots, 6$.
  - Sample with ten data points $\{2, 2, 1, 6, 2, 3, 3, 1, 4, 3\}$
  - Sample mean from raw data: $\langle x \rangle = (2 + 2 + 1 + 6 + 2 + 3 + 3 + 1 + 4 + 3)/10$
  - Data rearranged by outcome: $\{1, 1, 2, 2, 2, 3, 3, 3, 4, 6\}$
  - Bin counts n_1=2, n_2=3, n_3=3, n_4=1, n_5=0, n_6=1
  - Relative frequencies r_b: $r_1 = \frac{2}{10}, r_2 = \frac{3}{10}, r_3 = \frac{3}{10}, r_4 = \frac{1}{10}, r_5 = \frac{0}{10}, r_6 = \frac{1}{10}$.
  - We can find the same sample mean also via r_b:
    $\langle x \rangle = \frac{2}{10}1 + \frac{3}{10}2 + \frac{3}{10}3 + \frac{1}{10}4 + \frac{0}{10}5 + \frac{1}{10}6$.

- When, as in the above example, discrete data is binned by its elementary outcomes, sample means can therefore be calculated as

$$\langle x \rangle = \sum_{b=1}^{B} r_b\, x_b \qquad \text{with} \qquad r_b = n_b/N \tag{6.12}$$

In this case the answer exactly equals that obtained from $(1/N)\sum_{i=1}^{N} x_i$.

- When continuous data is binned or when discrete-data bins contain more than one elementary outcome, the relative frequency density is used, together with the bin midpoints x_b,

$$\langle x \rangle \simeq \sum_{b=1}^{B} \eta\, r_b\, x_b \qquad \text{with} \qquad r_b = n_b/N\eta. \tag{6.13}$$

- **Information loss**: The two ways of calculating of sample statistics directly from the data and from relative frequencies will not always yield exactly the same answers. In cases where (for discrete x_i) a bin contains more than one discrete outcome or (for continuous x_i) the bin ranges over a finite interval, the grouping of individual data points into bins inevitably results in the loss of the exact value of each x_i since it is replaced by the bin midpoint x_b. While the two methods do converge as the bin width η tends to zero, for true data processing this limit is not normally reached. Should the two methods yield different results, the calculation of a given sample statistic directly from the raw data is more reliable.

- Sample moments and in sample statistics in general can be approximated by relative frequencies in the same way,

$$\langle x^q \rangle \simeq \begin{cases} \sum_b r_b \, x_b^q & \text{for relative frequencies,} \\ \sum_b \eta \, r_b \, x_b^q & \text{for relative frequency densities,} \end{cases} \tag{6.14}$$

  and generally for any function of the data $h(x_i)$

$$\langle h(x) \rangle \;\simeq\; \sum_b r_b \, h(x_b) \quad \text{or} \quad \sum_b \eta \, r_b \, h(x_b). \tag{6.15}$$

6.5 Counts of functions of x_i

- As in probability theory, there are situations in which we want to consider statistics under transformations from the original raw data x_i to $u_i = \Phi(x_i)$, where Φ depends on the problem at hand and is known.

- Where Φ is bijective, the procedure for raw data is simple. Stay with discrete outcome data for the moment. Transforming each point x_i to u_i immediately permits raw-data statistics calculations. Where the transformed data is to be binned, one would first transform $\mathcal{A}(X)$ to $\mathcal{A}(U)$ then partition the latter into U-bins and determine relative frequencies in U-space,

$$r_b(U) \;=\; \frac{n_b = \text{Number of raw } \Phi(x_i) \text{ falling into } u\text{-bin b}}{N}$$

  and proceed to the desired sample statistics as before such as

$$\langle u \rangle = \frac{1}{N} \sum_{i=1}^{N} u_i \simeq \sum_{b=1}^{B} r_b \, u_b \qquad \langle h(u) \rangle = \frac{1}{N} \sum_{i=1}^{N} h(u_i) \simeq \sum_{b=1}^{B} r_b \, h(u_b) \qquad \text{etc}$$

- For data binned in x_i and the original relative frequency densities $r_b(X)$, the process of transformation to a relative frequency density $r_b(U)$ would involve a pseudo-Jacobian formed from the ratios of the respective bin widths. Details are beyond the scope of this course.

6.6 Relative frequency and probability

- It is crucial to recognise that **relative frequencies are not probabilities.** Even if we were to repeat a given experiment under identical conditions, the raw data $\mathcal{D}$ can and usually does change, and with it all the counts n_b, the relative frequencies r_b and all other sample statistics such as m_1, k_2 etc. By contrast, the functional form of the sampling probability (likelihood) $p(\mathcal{D} \,|\, \theta, \mathcal{I})$, never changes with changing data; only its value changes as different $\mathcal{D}$ is substituted into the same functional form for constant given parameters θ. There is a fundamental distinction between samples and their statistics on the one hand and the equivalent sampling probability. Data, counts and relative frequencies belong to the "Bottom" real world. In the sampling probability, the parameters θ are "Top" and unchanging, no matter which $\mathcal{D}$ we substitute into it.

- **Probability for frequentists**

  We briefly consider the traditional frequentist attempt to use relative frequencies to determine probabilities. In the frequentist approach, probability for a **discrete sampling space** is defined by means of an idealised limit: Given a dataset of N data resulting in n_b counts per bin $\mathcal{A}_b$, then the set of discrete probabilities is defined by frequentists as

$$P(X \in \mathcal{A}_b) = \lim_{N\to\infty} \frac{n_b}{N} = \lim_{N\to\infty} r_b.$$

- For continuous sampling spaces, the transition corresponds to the classic definition of a Riemann integral. Let the number of data points tend to infinity, $N \to \infty$; also let the number of bins become infinite while the bin width tends to zero, $B \to \infty$ and $\eta \to 0$. Then with $x_b = A + (b - \frac{1}{2})\eta$,

$$x_b \to x \qquad \eta \to dx \qquad r_b\,\eta \to p(x)\,dx \qquad \sum \to \int$$

  and we obtain a "histogram" with infinitely many bins of zero width which cover the original sampling space $\mathcal{A}$. The probability density then purportedly follows as the limit

$$P(X{=}x) = p(x) = \lim_{\substack{\eta\to 0\\ B\to\infty\\ N\to\infty}} \frac{n_b}{N\,\eta}. \tag{6.16}$$

- We emphasise that the above results and indeed the very approach are highly problematic and should not be given undue credence. The fact that convergence of the relative frequency to a fixed number often does occur is not a proof that it does so always. The proper framework to understanding probability is Bayesian *degree of belief* and the formal process of inference which it entails and which we shall introduce below.

- The only reason why these formulae still appear in this course is their continued use within Monte Carlo explorations of the Bayesian parameter spaces. For the moment, there seems to be no better way to do it, and we ourselves shall make use of such convergence arguments in Chapters 8 and 9.

6.7 Sample statistics and expectation values

- We are already familiar with sample statistics such as m_1, m_2, k_2 and sample averages in general. If and when one believes that r_b tends to the true probability, then the sample statistics would correspondingly tend to corresponding expectation values of the likelihood; for example

$$m_1 = \frac{1}{N}\sum_i x_i \quad \to \quad \mu_1 = \int dx\, p(x\,|\,\theta)\, x \tag{6.17}$$

  and so on.

- It must be emphasised that such convergences, if they happen at all, may happen only for large values of N. There is no general theorem regarding the rate of such convergence, and so, given a particular dataset $\mathcal{D}$ and its m_1, it is not clear what μ_1 is at all.

- The problem worsens if we realise that all such sampling statistics themselves fluctuate as we repeat an experiment to obtain many datasets.

- A telling example is that of gaussian-distributed real-x_i data. Probability theory tells us that the third cumulant is exactly zero, $\kappa_3 = \mu_3 - 3\mu_2\mu_1 + 2\mu_1^3 \equiv 0$, and yet the corresponding sampling cumulant $k_3 = m_3 - 3m_2m_1 + 2m_1^3$ is not zero at all, even if so-called bias corrections are introduced. All that can reasonably be said is that the set of such k_3 will have a histogram with a maximum near 0 and that this histogram would narrow as N increases.

- Given this context, it is very important to not confuse data averages with their theoretical counterparts, because they play very different roles in statistics: sample statistics such as m_1 are entirely data-dependent and independent of any theory, while expectation values such as μ_1 are entirely theory and have no data content. Never write $\langle\cdot\rangle$ when talking about an expectation value, and never write $E(\cdot)$ when you are talking about data. The fact that many books and physicists interchange the two notations only compounds the confusion.

- We end this section by mentioning an important formula resulting from this dubious approach, which nevertheless plays a significant role in much of experimental data analysis. Data is often displayed in terms of a *mean-and-standard-error* format, where $m_1 = \langle x\rangle$ is considered the best estimate of the true μ_1 and the standard deviation $\sqrt{k_2}$ of the sampling variance $k_2 = m_2 - m_1^2$ is used to provide a rough-and-ready estimate of the uncertainty associated with m_1. In order to take into account that the uncertainty pertains not to each individual x_i but to their average $\langle x\rangle$, the appropriate uncertainty is purportedly $\sigma_{\langle x\rangle} = \sigma_x/\sqrt{N}$, and the standard way of presenting experimental results is therefore

$$\langle x\rangle \ \pm \ \sqrt{\frac{\langle x^2\rangle - \langle x\rangle^2}{N}} \tag{6.18}$$

Chapter 7

Inference

There is no time in the hybrid course of 2019 for a proper exposition on inference. Very briefly, the proper way to do inference is based on Bayes' Theorem, which for the case of inferring the values of one or more parameters of interest $\boldsymbol{\theta}$ given some data $\mathcal{D}$ and under a model hypothesis $\mathcal{H}$ is

$$\boxed{p(\boldsymbol{\theta} \,|\, \mathcal{D}, \mathcal{H}) = \frac{p(\mathcal{D} \,|\, \boldsymbol{\theta}, \mathcal{H})\, p(\boldsymbol{\theta} \,|\, \mathcal{H})}{p(\mathcal{D} \,|\, \mathcal{H})} = \frac{p(\mathcal{D} \,|\, \boldsymbol{\theta}, \mathcal{H})\, p(\boldsymbol{\theta} \,|\, \mathcal{H})}{\int_{\mathcal{A}_{\boldsymbol{\theta}}} d\boldsymbol{\theta}\, p(\mathcal{D} \,|\, \boldsymbol{\theta}, \mathcal{H})\, p(\boldsymbol{\theta} \,|\, \mathcal{H})}.} \tag{7.1}$$

with

$p(\boldsymbol{\theta} \,|\, \mathcal{D}, \mathcal{H})$	=	posterior probability for $\boldsymbol{\theta}$ once the data has been acquired,
$p(\mathcal{D} \,|\, \boldsymbol{\theta}, \mathcal{H})$	=	the sample probability for the data $\mathcal{D}$, given knowledge of $\boldsymbol{\theta}$ (identical to the likelihood of $\boldsymbol{\theta}$),
$p(\boldsymbol{\theta} \,|\, \mathcal{H})$	=	the prior probability for the parameters, assigned by the observer based on the available information and hypothesis,
$p(\mathcal{D} \,|\, \mathcal{H})$	=	the "evidence" for $\mathcal{D}$ within hypothesis $\mathcal{H}$, an integral over the likelihood and prior; it plays a major role later.

For discrete x, the corresponding Bayes' Theorem would involve sums.

The simplest version of Bayesian parameter inference involving a Gaussian likelihood

$$p(x_i \,|\, \mu, \sigma) = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}},$$

logical independence, $p(\mathcal{D} \,|\, \mu, \sigma) \stackrel{\text{LI}}{=} \prod_i p(x_i \,|\, \mu, \sigma)$, and a uniform prior $p(\mu)$ results in a posterior which itself is also a Gaussian,

$$p(\mu \,|\, \mathcal{D}, \sigma) = \frac{\sqrt{N}}{\sigma}\, \frac{e^{-N(\mu-\langle x\rangle)^2/2\sigma^2}}{\sqrt{2\pi}} \tag{7.2}$$

which shows that the probability for the location parameter μ has a maximum where it equals sample average $m_1 = \langle x\rangle$ and that $E(\mu) = \langle x\rangle$, where $E(\mu) = \int d\mu\, \mu\, p(\mu \,|\, \mathcal{D})$. Importantly, this result also shows that the standard deviation for μ is found to be

$$\sigma_\mu = \sqrt{E(\mu^2) - E(\mu)^2} \;=\; \frac{\sigma}{\sqrt{N}}, \tag{7.3}$$

where $E()$ pertains to expectation values of the posterior, i.e. taking the data into account. This result validates the frequentist formula (6.18) without proving or assuming the latter's general validity. The parameter σ is considered fixed and known throughout.

Chapter 8

Introduction to Monte Carlo theory

8.1 Random Number Generators

Monte Carlo stands and falls with the generation of large sets of so-called random numbers. Uniformly-distributed random numbers $\{u_i\}_{i=1}^N$ are available on all computer systems in all numerical languages by means of a *random number generator* (RNG). They typically come in two versions, namely random positive integers distributed uniformly between 1 and some maximum integer, and floating-point numbers which are distributed uniformly over the open line interval $(0, 1)$ according to a uniform distribution

$$g(u') = 1, \qquad \mathcal{A}(U') = \{0 < u' < 1\}. \tag{8.1}$$

There are many different algorithms to generate these numbers. The science of doing so is highly sophisticated and arcane, making use of the full spectrum of inventiveness across computer science and number theory. While the details of RNG theory are beyond the scope of this course, the user is warned that bad RNG algorithms do exist and that it is up to him or her to exercise due diligence. Because errors and biases due to bad RNG algorithms are not easy to detect, they are doubly dangerous. Test suites for random number generators do exist. A basic precaution which should be exercised is to test whether, within the context of a given project, the use of two different RNG algorithms yields noticeably different results beyond normal statistical fluctuations.

Seeding: Any sequence of Random Number Generators starts with a so-called seed, which typically is an integer specified by the user. While the exact number used for the seed should be immaterial, the purpose of seeding the RNG with the same seed in successive runs is to obtain the exact sequence of "random" numbers for all the runs. This so-called *quenched randomness* allows exact reproducibility and is useful e.g. during program debugging.
Where quenching is not needed, seeds which differ from run to run are conveniently generated by using, for example, the system clock. Correspondingly, many modern RNGs select a seed automatically and silently, resulting in different number sequences for every run.

8.2 Generation of nonuniform random numbers: transformation method

Random numbers X which are distributed according to a non-uniform $p(x)$ can easily be obtained from $g(u')$ in cases where the transformation between U and X can be integrated and inverted analytically. The transformation method is based on the conservation of probability under transformation Eq. (4.20),

$$F_X(x) = \int_{-\infty}^{x} dx'\, p(x') = \int_{-\infty}^{\Phi(x)} du'\, g(u') = F_U(u)$$

with $d\Phi/dx > 0$ assumed, and since $g(u') = 1$ is uniform,

$$F_X(x) = F_U(u) = \Phi(x) = u. \tag{8.2}$$

Solving (8.2) for x,

$$x = F_X^{-1}(u) \tag{8.3}$$

yields, for every random number u_i which is generated from the uniform RNG, a number x_i such that

$$x_i \sim p(x)$$

where $\sim$ denotes "is distributed according to".

Example 1: x uniformly distributed between A and $A + L$, $p(x') = 1/L$; $\mathcal{A}(X') = (A, A + L)$:

$$u = F_X(x) = \int_{-\infty}^{x} p(x')\,dx' = \int_{A}^{x} p(x')\,dx' = \frac{x - A}{L} \tag{8.4}$$

$$x = Lu + A \equiv F_X^{-1}(u). \tag{8.5}$$

Example 2: Exponential distribution $p(x) = \alpha e^{-\alpha x}$, $\mathcal{A}(X) = [0, \infty)$ $\alpha > 0$:

$$u = F_X(x) = \int_{-\infty}^{x} p(x')\,dx' = 1 - e^{-\alpha x} \tag{8.6}$$

$$-\alpha x = \ln(1 - u) \tag{8.7}$$

$$x = -\frac{1}{\alpha}\ln(1 - u). \tag{8.8}$$

The set of distributions for which the transformation method works is limited. While there are other methods to generate random numbers, for example by the so-called *rejection method*, we do not have time to treat these.

8.3 Monte Carlo "hit-or-miss" integration

Almost all theory of Monte Carlo can be couched in terms of approximating solutions to integrals and to estimating expectation values. Integrals and expectation values are closely related, but expectation values are more general because they may be taken with respect to an arbitrary probability density, while integrals implicitly assume that every part of the volume to be integrated has the same weight, and that assumption implies use of a uniform probability.

There are, of course, many methods to solve integrals numerically, and Monte Carlo is often not the best method. Nevertheless, there are situations where conventional numerical methods do not apply or are less efficient in terms of CPU time and accuracy such as complicated integral boundaries or integrands and high-dimensional integrals.[1]

Throughout this section, We shall make use of the following simple one-dimensional example and then generalise to higher dimensions. Given a known function $h(x)$, the example involves seeking a numerical approximation to the integral

$$I_h = \int_0^L dx\, h(x). \tag{8.9}$$

[1] If an integral can be solved analytically, it is of course unnecessary to use Monte Carlo methods; on the other hand, an analytical answer may be used to test the efficiency and accuracy of a given MC method.

as well as a numerical approximation to the expectation value of $h(x)$. For simplicity, the domain of our example integration is the finite interval $\mathcal{A}(X) = (0, L)$; more advanced Monte Carlo methods can handle infinite intervals. Assume further that $h(x)$ is *bounded and positive*,

$$0 \leq h(x) \leq H \qquad \forall\, x \in \mathcal{A}(X), \tag{8.10}$$

but in general h does not have to be positive.

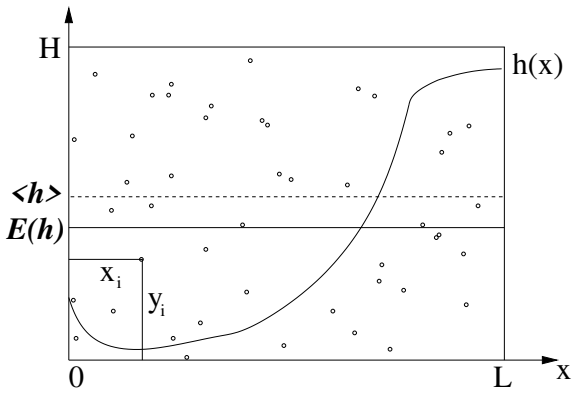


Fig 1: Hit-or-miss Monte Carlo. Point pairs (x_i, y_i) are generated randomly with x_i distributed uniformly in the sampling space $\mathcal{A}(X) = (0, L)$ of $h(x)$, while y_i is distributed uniformly between the minimum and maximum values that $h(x)$ attains, $\mathcal{A}(Y) = (0, H)$. Points (x_i, y_i) which fall *under* the corresponding $h(x_i)$ are "hits", while points which lie *above* h are "misses". The sample average $\langle h \rangle$ (dashed line) estimates the expectation value $E(h)$ (solid line) which is related to the integral by $I_h = L\, E(h)$.

The simplest numerical approach, called the "hit-or-miss" method, aims to determine the area under the curve directly. Look at Fig. 1: The area can be estimated by generating points, distributed randomly in the rectangle between $(0, 0)$ and (L, H), and counting how many of these fall within the area under $h(x)$ ("hits") and how many in the area above it ("misses"). Specifically, a set of N pairs of random variables (x_i, y_i) is generated where x_i is uniformly distributed in $(0, L)$ and y_i is uniformly distributed in $(0, H)$. If n is the number of hits and $N - n$ the number of misses according to

$$y_i < h(x_i) \Rightarrow \text{"hit"} \qquad\qquad y_i > h(x_i) \Rightarrow \text{"miss"} \tag{8.11}$$

then the true value I_h of the integral is estimated by the "estimator" $\hat{I}_{HM}$,

$$\frac{\hat{I}_{HM}}{LH} = \frac{n}{N}\,. \tag{8.12}$$

The accuracy of $\hat{I}_{HM}$ will clearly improve with increasing N; indeed, the uncertainty of $\hat{I}_{HM}$ is given by the standard deviation of the sample mean whose accuracy improves as $1/\sqrt{N}$. It is, however, also quite clear that the approach is not very efficient since a substantial fraction of random numbers are generated "in vain" since they are thrown away eventually. This aspect will be investigated further in Section 8.6.

It is also clear that HM Monte Carlo is robust; it can be used for highly complicated problems with many preconditions and constraints. The calculation merely needs to test whether a given random point satisfies each of those conditions, in which case it counts as a "hit". This can be couched in mathematical language for integrals of any dimension with vector variables $\boldsymbol{x}$ and any function $0 \leqslant h(\boldsymbol{x}) \leqslant H$ which is to be integrated. The hit-or-miss algorithm then requires us to generate

randomly uniformly-distributed points $\boldsymbol{x}_i \sim p_u(\boldsymbol{x})$ as well as y_i uniformly distributed over $[0, H]$. Using the theta function $\Theta(C)$, a "hit" for the i-th random point $\boldsymbol{x}_i$ is defined by

$$n_i = \Theta[h(\boldsymbol{x}_i) - y_i].$$

The estimator for the integral is then

$$\frac{\hat{I}_h}{|\mathcal{A}(\boldsymbol{x})|} = \frac{\sum_{i=1}^N n_i}{N} \tag{8.13}$$

While the use of MC to approximate an integral is helpful for visualisation, it is more generally used in the context of approximating expectation values. For those purposes, we define for the above one-dimensional example a uniform probability

$$p_u(x) = \frac{1}{L}, \qquad \mathcal{A}(X) = (0, L), \tag{8.14}$$

where from now on u simply indicates "uniformity" rather than a random variable. The expectation value of a function $h(x)$ within p_u is then proportional to the integral,

$$E[h(x)\,|\,p_u(x)] = E_u(h) = \frac{1}{L}\int_0^L dx\, h(x) = \frac{I_h}{L}. \tag{8.15}$$

and the estimator for $E_u(h)$ for the Hit-or-miss method in the above one-dimensional example is

$$\langle h\rangle_{u,HM} = \hat{E}_u(h) = \frac{\hat{I}_h}{L} = H\frac{n}{N} \tag{8.16}$$

and in general, for higher-dimensional spaces with variables $\boldsymbol{x}$ and uniform probability

$$p_u(\boldsymbol{x}) = \frac{1}{|\mathcal{A}(\boldsymbol{X})|} \qquad \forall\, \boldsymbol{x} \in \mathcal{A}(\boldsymbol{X}) \tag{8.17}$$

a true expectation value

$$E_u[h(\boldsymbol{x})] = \int_{\mathcal{A}(\boldsymbol{X})} d\boldsymbol{x}\, h(\boldsymbol{x})\, p_u(\boldsymbol{x}) \;=\; \frac{1}{|\mathcal{A}(\boldsymbol{X})|}\int_{\mathcal{A}(\boldsymbol{X})} d\boldsymbol{x}\, h(\boldsymbol{x}) \;=\; \frac{I_h}{|\mathcal{A}(\boldsymbol{X})|} \tag{8.18}$$

the Hit-or-miss estimator will be

$$\langle h\rangle_{u,HM} = H\frac{n}{N} = \frac{H}{N}\sum_{i=1}^N \Theta[h(\boldsymbol{x}_i) - y_i] \tag{8.19}$$

which can also be written as $\hat{I}_h/|\mathcal{A}(\boldsymbol{X})|$.

8.4 Lattice Sampling and Simple Sampling

In order to motivate stochastic methods, we first consider "lattice sampling" (LS) or "grid sampling" which is deterministic, primitive and robust: the support, domain, or whatever the basis of $h(x)$ is called is covered by a lattice or grid consisting of N lattice points. Integration is then the limit of the sum over the areas of N rectangles of fixed width $\Delta x = \eta = L/N$ and heights $h(x_b)$ at either the single-sample "bin" midpoints or, for that matter, the left bin boundaries $x_i = (i - \frac{1}{2})\eta = (i - \frac{1}{2})L/N$,

$$I_h \;=\; \lim_{\substack{\eta\to 0\\ N\to\infty}} \hat{I}_{LS} \tag{8.20}$$

$$\text{with} \quad \hat{I}_{LS} \;=\; \sum_{i=1}^N \eta\, h(x_i) = \frac{L}{N}\sum_{i=1}^N h(x_i). \tag{8.21}$$

if the single-sample "bins" all have the same width. The corresponding expectation-value of $h(x)$ for the uniform probability (8.14),

$$E_u(h) \equiv E[h(x) \,|\, p_u(x)] = \int_0^L dx\, h(x)\, p_u(x) \tag{8.22}$$

is estimated by the sample mean

$$\langle h \rangle_{u,LS} = \frac{1}{N} \sum_{i=1}^{N} h(x_i) \qquad x_i = (i - \tfrac{1}{2})N. \tag{8.23}$$

Note that in LS the number of bins B and the number of particles N is the same.

In *Simple Sampling* (SS), also known as *Crude Monte Carlo* = CMC, we similarly sum over many rectangles with heights $h(x_i)$. The rectangle left lower corners x_i are generated from $p_u(x)$, resulting in random rectangle widths ε_i between neighbouring points. The method is called *Simple Sampling* because points x_i have the same chance to appear in the sum, irrespective of their location in $\mathcal{A}(X)$ and of the magnitude of $h(x_i)$ at that point.

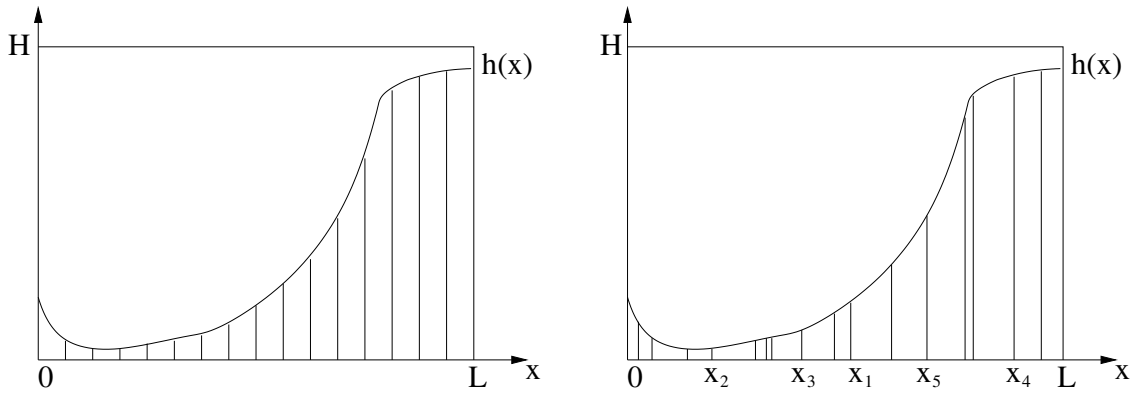


Fig 2: Left: The lattice-sampling integral, based on evaluations of the function $h(x_i)$ at equally spaced points x_i. Right: Evaluation of the same integral by means of a set of uniformly distributed random numbers $\{x_i\}$ in the Simple Sampling method.

Fig. 2 illustrates the difference between the lattice-based integral $\hat{I}_{LS}$ and the Simple Sampling integral

$$\hat{I}_{SS} = \sum_{i=1}^{N} \varepsilon_i h(x_i) \tag{8.24}$$

where ε_i is the distance between i-th random number x_i and its nearest neighbour to the left. Correspondingly, for x_i sampled from the uniform probability $p_u(x) = 1/L$ the sample mean for the raw simdata is

$$\langle h \rangle_{u,SS} = \frac{1}{N} \sum_{i=1}^{N} h(x_i) \qquad x_i \sim p_u(x). \tag{8.25}$$

which is also an estimator for $E_u(h)$.

In order to check whether such estimators are at least consistent with the true integral or true expectation values, we now take a frequentist limit, namely that of letting the number of simdata points N grow to infinity. We know, of course, that this is conceptually wrong and that a better method would involve a fully Bayesian analysis. However, since most of the Monte Carlo literature hardly

understands the probabilistic roots of the algorithms, never mind the differences between frequentist and Bayesian concepts, we will for the present follow the primitive route of testing for convergence under an unattainable limit.

The standard (frequentist) convergence test is based on the dual representation of sample means set out in Section 6.4 and specifically the two ways of calculating sample means. In Eq. (6.11), the sample average $\langle h(x)\rangle = \frac{1}{N}\sum_i h(x_i)$ was taken over the raw data, while in Eq. (6.15), the sample average was approximated as

$$\langle h\rangle_u = \sum_{b=1}^{B} \eta\, h(x_b)\, r_b, \tag{8.26}$$

where x_b is the bin midpoint as before, $\eta = L/B$ is the bin width and r_b are the relative frequency densities

$$r_b \;=\; \frac{n_b}{\eta\, N} = \frac{(\text{number of } x_i\text{'s in bin } b)}{\eta \cdot (\text{total number } x_i\text{'s})} \tag{8.27}$$

The convergence argument runs that r_b is an experimental quantity which mostly (but not always) converges to the probability density $p(x)$ in the set of limits $\eta \to 0, B \to \infty, N \to \infty$ so that $\langle h\rangle_u$ converges to $E_u(h)$ since, in terms of Eq. (8.26) and the usual limits $\sum_b \eta \to \int dx$ and $r_b \to p_u(x)$

$$\lim_{N\to\infty} \langle h\rangle_u = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} h(x_i) = \lim_{\substack{N\to\infty \\ \eta\to 0 \\ B\to\infty}} \sum_{b=1}^{B} \eta\, h(x_b)\, r_b = \int_0^L dx\, h(x)\, p_u(x) = E_u(h). \tag{8.28}$$

A similar convergence test for the integral estimator (8.24) runs as follows. As N increases, the individual ε_i's will tend to be distributed around a mean $\langle\varepsilon\rangle = (1/N)\sum_{i=1}^{N}\varepsilon_i = L/N$, so that to good approximation, with $r_b \to p_u(x) = 1/L$

$$\hat{I}_{SS} = N\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i h(x_i) \simeq N\sum_b \eta\, r_b\, [\langle\varepsilon\rangle\, h(x_b)] = L\sum_b \eta\, r_b\, h(x_b) = \int_0^L dx\, h(x). \tag{8.29}$$

Note that in all the above cases the distinction between relative frequency r_b and the probability density $p(x)$ must be strictly maintained: convergence may well occur in the limit, but in practice that limit is never attained exactly.

In spaces with K dimensions, the uniform probability used in Simple Sampling would be the $p(\boldsymbol{x})$ of Eq. (8.17) i.e. the inverse of the size of the sampling space of $\boldsymbol{x}$. Correspondingly, there would be K different binnings in each direction with the K-dimensional boxes of widths $\boldsymbol{\eta}$ converging to K-dimensional integrals.

8.5 Importance Sampling

8.5.1 Basic Idea

It can be shown that Simple Sampling is more efficient than hit-or-miss. The strategy of evaluating $h(x)$ at N points $x_i \sim p_u$ uniformly distributed over $\mathcal{A}(X)$ is, however, obviously not optimal because points x_i where $h(x_i)$ is small (as e.g. in the left part of h in Fig. 2) contribute little to the mean (8.29), while points x_i where $h(x_i)$ is large contribute more. Both nevertheless cost the same amount of CPU time. The problem is actually worse than that. In cases where $h(x)$ is extremely peaked, as commonly occurs when we evaluate the likelihood in parameter space, Simple Sampling will result in huge contributions being made by that small number of points x_i which happen to be under the peak,

while the contribution of the large majority of points is negligible. In such a dire situation, one can never be certain that the estimate obtained is stable.

It will clearly be more effective to generate a set of random numbers $\{x_i\} \sim f$ which are *not* uniformly distributed but rather follow a probability density $f(x)$ which resembles $h(x)$ in shape as closely as possible, yielding more numbers x_i in those subregions of $\mathcal{A}$ where h is large.

This idea is realised by *importance sampling*. Given a function $h(x)$ which is to be integrated, seek a probability density $f(x)$ with the following properties:

- It must have the same sampling space as h (if necessary, transformation methods of Section 8.2 can be employed to "stretch or squeeze" the outcome space),
- $f(x)$ may not be zero at any point $x \in \mathcal{A}$,
- there must be a way to generate samples which are distributed according to $f(x)$ according to one of the known methods,
- the *shape* of $f(x)$ should resemble that of $h(x)$ as closely as possible. Given that f is a probability density, the integrals of f and h will necessarily differ by a factor I_h since

$$\int_{\mathcal{A}} dx\, f(x) = 1 \qquad\qquad \int_{\mathcal{A}} dx\, h(x) = I_h. \tag{8.30}$$

  For this reason, we can talk about approximation only in shape but not in magnitude.

- It is *not* necessary that $f(x) > h(x) \quad \forall x \in \mathcal{A}$ as in the case of the rejection method.

In practice, it may not be easy to find an f which is both usable for x_i generation as well as close in shape to h. Nevertheless, any f which resembles h more than a uniform $1/L$ shape will be better.

The progression from Simple Sampling to Importance Sampling is represented in Fig. 3. The two panels in the left hand column represent the Simple Sampling of Eq. (8.21), nonuniform h sampled by uniform f, while those in the right hand column represent the Importance Sampling of Eq. (**??**), almost uniform h/f sampled with nonuniform f. The vertical lines denote a few of the many randomly generated values x_i. The density of these lines are determined directly by the pdf $p_u(x)$ and $f(x)$, shown in the lower two panels, according to which the x_i were generated. Clearly, in Importance Sampling the density of sampling points x_i is higher in those parts where $h(x)$ is larger, while in Simple Sampling the regions of small $h(x)$ receive equal but unnecessary attention.

We now consider the resulting integral and sample mean estimators. First we generate a set of f-distributed random numbers $\{x_i\}$. Of course this set cannot be used directly to evaluate $\hat{I}_{IS}$ because different subregions of $h(x)$ would not be equally represented in this sample. The estimator

$$\hat{I}_{IS} = \sum_i \varepsilon_i\, h(x_i) \qquad x_i \sim f(x) \tag{8.31}$$

is of course correct and efficient (in the sense that we evaluate h where it matters), but it is computationally expensive to find the ε_i even in one dimension, while in higher dimensions it becomes very hard.

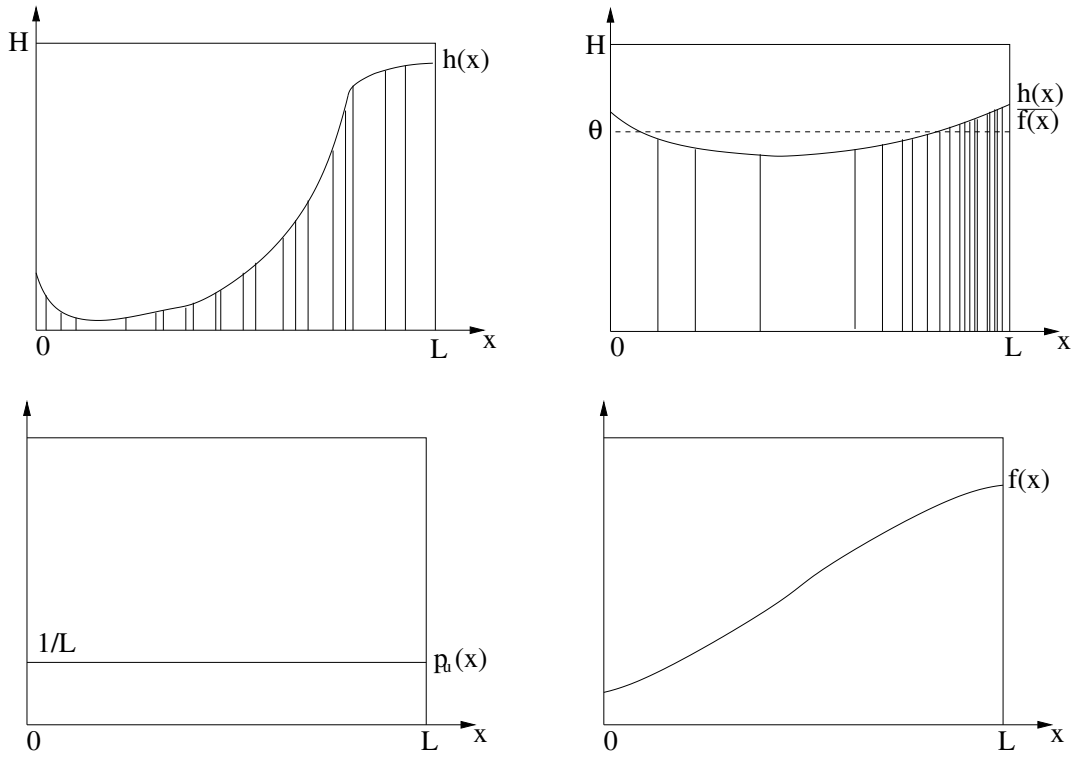


Fig 3: Graphical comparison of Simple Sampling (left) and Importance Sampling (right).

Unlike the Simple Sampling case, the convergence test in this case cannot rely on $\varepsilon_i \to \langle\varepsilon\rangle = L/N$ because the x_i are not uniformly distributed. Rather, on considering large N, small bins and comparatively smooth $f(x)$, we must argue that the typical size of the nearest-neighbour ε_i in the bin b is related to the inverse of the probability density in that bin,

$$\varepsilon_i \to \langle\varepsilon\rangle_b = \frac{1/f(x_b)}{N} \tag{8.32}$$

so that, with $r_b \to f(x)$, $(1/N)\sum_i \to \sum_b$,

$$\lim_{\substack{\eta\to 0\\ B\to\infty\\ N\to\infty}} \hat{I}_{IS} = N\frac{1}{N}\sum_{i=1}^{N} \varepsilon_i\, h(x_i) \qquad x_i \sim f(x) \tag{8.33}$$

$$= \lim_{\substack{\eta\to 0\\ B\to\infty\\ N\to\infty}} N\sum_b \eta\, r_b\, \langle\varepsilon\rangle_b\, h(x_b) = N\int_{\mathcal{A}} dx\, f(x)\,\frac{1/f(x)}{N}\,h(x) \tag{8.34}$$

$$= \int_{\mathcal{A}} dx\, h(x) = I_h. \tag{8.35}$$

Importance sampling of an expectation value is both simpler and more general. In the simplest case, we would want to find $E_p(h) = E[h(x)\,|\,p(x)]$ for some nonuniform probability $p(x)$ which both resembles h in shape and from which we can sample. In that case, we may simply evaluate

$$\langle h\rangle_{p,IS} = \frac{1}{N}\sum_i h(x_i) \qquad x_i \sim p(x) \tag{8.36}$$

because in the limit, with $r_b \to p(x)$,

$$\lim_{N\to\infty}\frac{1}{N}\sum_i h(x_i) = \lim_{\substack{\eta\to 0\\ B\to\infty\\ N\to\infty}} \sum_b \eta\, r_b\, h(x_b) = \int dx\, p(x)\, h(x). \tag{8.37}$$

More generally, however, we either cannot sample directly from $p(x)$ or else it differs in shape from $h(x)$. In that case, we must sample from that $f(x)$ which does resemble h and which can be sampled, but compensate by sampling not $h(x_i)$ alone but weight it by the ratio p/f,

$$\langle h\rangle_{p,IS} = \frac{1}{N}\sum_i h(x_i)\,\frac{p(x_i)}{f(x_i)} \qquad x_i \sim f(x), \tag{8.38}$$

because this converges to the correct expectation value (with $r_b \to f(x)$, not $p(x)$),

$$\lim_{N\to\infty}\frac{1}{N}\sum_i \frac{h(x_i)\,p(x_i)}{f(x_i)} = \lim_{\substack{\eta\to 0\\ B\to\infty\\ N\to\infty}}\sum_b \eta\, r_b\,\frac{h(x_b)\,p(x_b)}{f(x_b)} = \int_{\mathcal{A}} dx\, f(x)\,\frac{h(x)\,p(x)}{f(x)} = E_p(h). \tag{8.39}$$

8.5.2 Importance Sampling for general expectation values

We can now generalise. The goal is to obtain a numerical estimate for the expectation value of a function $h(\boldsymbol{x})$ in some K-dimensional space $\mathcal{A}(\boldsymbol{X})$ with respect to a probability density $p(\boldsymbol{x})$, not necessarily uniform. When $p(\boldsymbol{x})$ can be sampled directly, the sample mean of h is simply

$$\boxed{\langle h\rangle_{p,IS} = \frac{1}{N}\sum_{i=1}^{N} h(\boldsymbol{x}_i) \qquad \text{with} \quad \boldsymbol{x}_i \sim p(\boldsymbol{x})} \tag{8.40}$$

because, by the same chain of argument as above, this converges to $\int d\boldsymbol{x}\, h(\boldsymbol{x})\, p(\boldsymbol{x}) = E_p(h)$. If, on the other hand, $p(\boldsymbol{x})$ cannot be sampled directly, we somehow guess or find a probability density $f(\boldsymbol{x})$ which resembles either $h(\boldsymbol{x})$ in shape, or $p(\boldsymbol{x})$, or optimally the product $h(\boldsymbol{x})p(\boldsymbol{x})$ itself. We generate simdata $\{\boldsymbol{x}_i\}$ from f and then estimate the sample mean from the generalisation of (8.38),

$$\boxed{\langle h\rangle_{p,IS} = \frac{1}{N}\sum_{i=1}^{N} \frac{h(\boldsymbol{x}_i)\,p(\boldsymbol{x}_i)}{f(\boldsymbol{x}_i)} \qquad \text{with} \quad \boldsymbol{x}_i \sim f(\boldsymbol{x})} \tag{8.41}$$

and by the same methods employed above, it can be shown that

$$\lim_{N\to\infty}\langle h\rangle_{p,IS} = E[h(\boldsymbol{x})\,|\,p(\boldsymbol{x})] = E_p(h). \tag{8.42}$$

Finally, the corresponding Importance Sampling method of estimating integrals in high-dimensional would be

$$\boxed{\hat{I}_{IS} = \sum_{i=1}^{N} \boldsymbol{\varepsilon}_i\, h(\boldsymbol{x}_i) \qquad \text{with} \quad \boldsymbol{x}_i \sim f(\boldsymbol{x})} \tag{8.43}$$

but this hides the problems associated with the practicalities of using nearest-neighbour intervals in high-dimensional spaces.

8.5.3 Algorithmic probabilities

Monte Carlo estimation of expectation values and integrals in general is necessary because such integrals can often not be calculated analytically. An important special case is that where the probability $p(\boldsymbol{x})$ itself is known only algorithmically, i.e. no closed functional form exists. All we have is some positive function $p' : \boldsymbol{x} \to p'(\boldsymbol{x})$ which is known to represent a probability and which can be evaluated uniquely at all $\boldsymbol{x} \in \mathcal{A}(\boldsymbol{X})$. In this case the normalised probability $p(\boldsymbol{x})$ above must be equated to $p'(\boldsymbol{x})$ divided by its integral,

$$p(\boldsymbol{x}) = \frac{p'(\boldsymbol{x})}{\int_{\mathcal{A}(\boldsymbol{X})} d\boldsymbol{x}\, p'(\boldsymbol{x})} \tag{8.44}$$

where the integral, too, cannot be evaluated analyticall but must be estimated. Any expectation value of $h(\boldsymbol{x})$ with respect to $p(\boldsymbol{x})$ must therefore be written as

$$E_p(h) = \frac{\int d\boldsymbol{x}\, h(\boldsymbol{x})\, p'(\boldsymbol{x})}{\int d\boldsymbol{x}\, p'(\boldsymbol{x})} \tag{8.45}$$

and both numerator and denominator must be estimated numerically by

$$\langle h \rangle_{p',IS} = \frac{1}{N} \sum_{i=1}^{N} \frac{h(\boldsymbol{x}_i)\, p'(\boldsymbol{x}_i)}{f(\boldsymbol{x}_i)} \qquad \boldsymbol{x}_i \sim f(\boldsymbol{x}) \tag{8.46}$$

$$\langle 1 \rangle_{p',IS} = \frac{1}{N'} \sum_{j=1}^{N'} \frac{p'(\boldsymbol{x}_j)}{f(\boldsymbol{x}_j)} \qquad \boldsymbol{x}_j \sim f'(\boldsymbol{x}) \tag{8.47}$$

where in principle f and f' can be different sampling probabilities and where in principle the indices i and j are different, i.e. the simdata $\{\boldsymbol{x}_i\}$ and $\{\boldsymbol{x}_j\}$ are independent. Naturally, doubling the simdata to $2N$ comes at the cost of doubling the CPU time. Many implementations therefore choose to save CPU time at the cost of introducing a degree of non-independence: they set $N' = N$, $f' = f$ and use the same simdata $\{\boldsymbol{x}_i\}$ both in $\langle h \rangle_p$ and $\langle 1 \rangle_p$, resulting in the Importance Sampling prescription

$$\langle h \rangle_{p,IS} = \frac{\displaystyle\sum_i \frac{h(\boldsymbol{x}_i)\, p'(\boldsymbol{x}_i)}{f(\boldsymbol{x}_i)}}{\displaystyle\sum_i \frac{p'(\boldsymbol{x}_i)}{f(\boldsymbol{x}_i)}} \qquad \boldsymbol{x}_i \sim f(\boldsymbol{x}) \tag{8.48}$$

8.5.4 Limitations and problems of Importance Sampling

Importance Sampling is a clear improvement over Simple Sampling. However, there are two problems, both in relation to sampling higher-dimensional spaces.

(a) **Convergence tests:** The convergence tests used extensively in the derivations leave much to be desired. They merely prove that the finite-N formula is not inconsistent with the asymptotic result but give no general indication of the accuracy or stability far from that asymptote. They also provide little guidance as to the number of samples N needed to achieve a desired accuracy. The only rough guide we have is that N must increase substantially with K, the dimensionality of $\mathcal{A}(X)$.

(b) **Finding f:** In one, two and even three dimensions, it is possible to visualise the location and shape of the function $h(x)$, so that the choice and design of $f(x)$ becomes a matter of inspection and setting parameters of $f(x)$. When more than three dimensions are involved, however, it is often not possible to know the shape or location of $h(x)$, so that the design of $f(x)$ becomes hard or impossible.

(c) **Dividing by near-zero:** It is hard to prevent the ratio $h(x_i)/f(x_i)$ from becoming very small and very large, depending on the shapes of h and f. This can have bad or even catastrophic consequences. For example, if h/f is very large in a region where h itself is small, simply because f is so much smaller in that region, then that region can with a single large contribution skew or even overwhelm contributions from other regions. The bottom line is that statistics of ratios of random variables is a dangerous game.

8.6 Efficiency and variance of estimators

8.6.1 General picture

Given the different Monte Carlo methods to calculate integrals and sample means, what criterion should we use to choose one method over another?

The speed of computation naturally depends strongly both on the hardware configuration of the computer system and on the specific code that is being run. A sensible comparison of Monte Carlo methods must therefore be based on a ratio of speeds: If Method 1 requires a time T_1 and Method 2 a time T_2 for the same calculation and coding language, the relative efficiency of the methods is given by T_1/T_2.

Secondly, the quest for accuracy leads the researcher directly to compare the *variances* of the two methods. The relative *efficiency* of Method 1 compared to Method 2 is hence quantified by a double ratio

$$V = \frac{T_2 \ \mathrm{var}(\hat{I}_2)}{T_1 \ \mathrm{var}(\hat{I}_1)} \tag{8.49}$$

A small V implies that Method 2 is better than Method 1. Here, the variance can as usual have one of two meanings: sometimes it is possible to calculate theoretical variances κ_2, while sample variances k_2 may have to be used in other cases.

As shown before, the uncertainty $\sigma(\hat{I}_h) = [\kappa_2(\hat{I}_h)]^{1/2}$ of results obtained by Monte Carlo methods decreases as $1/\sqrt{N}$. This implies that highly accurate Monta Carlo solutions become very expensive in CPU time. Nevertheless, Monte Carlo will be more efficient than numerical approaches in cases where the integrand or the sampling space is complicated or if algorithmic conditions, rules, ifs and buts apply.
Secondly, Monte Carlo methods become more competitive when *high-dimensional* integrals must be calculated since the above theoretical variances can without further approximation be written in terms of such higher-dimensional spaces. The uncertainty $\sigma(\hat{I}_h)$ has the same $1/\sqrt{N}$ dependence irrespective of the dimension of the integral.

It needs to be emphasised that *the quality of Monte Carlo results depend on the quality of the random number generator.* Program code may be great, results may look attractive, and even the calculated variance may be small. Nonetheless, the results will be completely wrong if the RNG has an inherent *bias* and/or the sampling space is not sampled properly, and/or the sample $\{x_i\}$ is not exactly distributed according to the desired f. The error resulting from bias of the RNG will, as we know, grow with N. Critical and wide-ranging testing of a random number generator is therefore not optional but vital. The same RNG must even be tested again for different applications.

Chapter 9

The Ising Model and Markov Chain Monte Carlo

The notes in this section do not replace the books but are complementary to them. You should read at least those sections in the books which are relevant to the basic work.

9.1 Physics background: experimental facts

Most atoms and molecules have a magnetic dipole which is mostly the result of the spin of one or more of the electrons in the outer shells. Depending on the interaction between the dipoles of single atoms, the following generic behaviour can result:

- Matter is characterised as a **paramagnet** if its internal dipoles interact only with an external magnetic field and not with other internal dipoles.
- In the case of a **ferromagnet**, there is a "positive" interaction between its dipoles or spins in the sense that they prefer to point in the same direction. This results in **magnetisation**.
- In a **antiferromagnet**, on the other hand, the interaction between dipoles is negative so that they prefer to point in directions opposite to that of their neighbouring spins ("anti-alignment").
- It is an experimental fact that every ferromagnet has a **Curie temperature** T_c: for temperatures smaller than T_c, the material is magnetised, while above T_c, the net magnetisation is zero.
- While iron is a ferromagnet, it is nevertheless mostly unmagnetised. The reason is that the macroscopic material is divided in to many "microdomains". While within each microdomain the atoms are arranged in an orderly lattice and thus form a collective micromagnet, the different domains have magnetisations in many different directions so that the overall magnetisation averages out to zero.

9.2 Basics of the Ising Model

- The Ising model is a *model* in the sense that it only captures the essentials of local interaction and ignores many of the complications which occur in physical magnetic systems. Nonetheless, it exhibits complicated behaviour and is therefore a valuable laboratory to study cause and effect of certain microscopic interactions.
- The Ising model is a toy model for a single microdomain. Normally, the domain is modelled as a **square lattice** in one, two or three dimension. Other structures such as hexagonal lattices are also possible.

- The **One-dimensional Ising model** can be solved analytically, see Schroeder pp. 341ff, but the solution gives no help for the two-dimensional case.
- We normally use the two-dimensional lattice because the one-dimensional lattice exhibits only trivial behaviour, while the computation of the three-dimensional lattice is slow. We define the side length of the lattice in terms of L units with an assumed lattice constant of 1. Therefore, we have $N = L^2$ **lattice points** in the two-dimensional case.
- A lattice point would normally be identified by two indices (i, j) each with values $1, 2, \ldots, L$. For both analytical and computational work it is, however, often easier to use only a single index $i = 1, 2, \ldots, N{=}L^2$. In the C programming environment, these indices would of course change to $i = 0, 1, 2, \ldots, (N{-}1)$.
- In the Ising model, each lattice spin has only **two possible states** $s_i = +1$ or $s_i = -1$. These spins are therefore dimensionless, i.e. we have automatically divided by the Bohr magneton μ_0.
- A *microstate* r in the Ising model is a particular configuration of the entire lattice, i.e. the state of each of the N spins is specified,

$$r \equiv \{s_1, s_2, \ldots, s_N\} \tag{9.1}$$

- The Ising model assumes that interactions occur only between **nearest neighbours**, abbreviated "n.n.": a given spin is influenced by the four spins immediately left, right, above and below it.
- The total energy of interaction is[1]

$$E_r = H = -J \sum_{\text{n.n. } i,j} s_i s_j \tag{9.2}$$

  where J is a positive constant and n.n. represents the full sum of $i, j = 1, \ldots, N$ but limited to only those cases where i and j are nearest neighbours. The minus sign ensures that spins pointing in the same direction have a lower interaction energy than spins pointing in opposite directions.
- Note: **Negative energies** are not a problem: we could add a large positive constant E_{large} in both numerator and denominator of all probabililities without changing the calculation in any way. All calculations eventually relate to energy differences and not to absolute energies.
- There are, of course, many microstates with the same energy.
- You should be aware of possible **"double counting"** of energies ε_{ij} of the interaction between a particular pairs (i, j).
- In the presence of an external magnetic field B, the energy of a microstate changes to

$$E_r = H = -J \sum_{\text{n.n. } i,j} s_i s_j \; - B \sum_i s_i \,. \tag{9.3}$$

  For simulation purposes, we again usually work with dimensionless energy E/J.
- For a particular microstate $r = \{s_1, \ldots, s_N\}$, the **magnetisation** is simply the sum of all spins' values,

$$M_r = \sum_i s_i \,. \tag{9.4}$$

[1] The notation H implies that we really here are dealing with a **Hamiltonian**. From the thermodynamic point of view, H is the internal energy.

- The Ising model is symmetric under the operation "flip all spins simultaneously", $s_i \to -s_i, i = 1, \ldots, N$. In other words, for each state $r = \{s_1, \ldots, s_N\}$ with magnetisation M_r there exists a symmetric opposite state $r' = \{-s_1, \ldots, -s_N\}$ with magnetisation $-M_r$ but exactly the same energy, $E_{r'} = E_r$. For some calculations, it is therefore necessary to use the absolute value

$$|M_r| = \Big|\sum_i s_i\Big|. \tag{9.5}$$

- The **size of the lattice** L will play a critical role in calculations. The general rule for the lattice size is "the larger the better", but a larger lattice comes at the cost of more CPU time. To diminish the influence of nonstandard spins on the boundaries of the lattice with only two or three nearest neighbours, the common practice is to use **periodic boundary conditions**, where a spin on the edge is given an interaction with a "neighbour" on the other side of the lattice, i.e. spins $s_{L,j}$ interact with $s_{1,j}$ and spins $s_{i,L}$ interact with $s_{i,1}$ etc.

  Newman and Barkema, however, use **helical boundary conditions** because the program code runs faster. Carefully study their subroutine which will be supplied to you.

9.3 Statistical physics of the Ising model

There are different ways of defining systems in statistical physics. Three so-called "ensembles" are relevant for our purposes: (a) the case where we place no limitation on the occurrence of any possible microstate, (b) an isolated system where only such microstates are permitted which have a total energy equal to a fixed constant E (normally called the **microcanonical ensemble**), and (c) a system where the total energy may change but the temperature is kept constant, the so-called **canonical ensemble**.

1. **Simple sampling: No heat bath, no limitation on energy:** In this case, the question being asked is: "In what microstate is the system?" and the answers to this are described by the variable r. For N spins each with two possible states $\mathcal{A}(s_i) = \{-1, +1\}$, the total sampling space is $\mathcal{A}_r = \mathcal{A}(s_1, \ldots, s_N) = \mathcal{A}(s_1) \otimes \cdots \otimes \mathcal{A}(s_N)$ and so there are

$$|\mathcal{A}(\{s_i\})| = \Omega = \sum_{s_1=-1}^{+1} \cdots \sum_{s_N=-1}^{+1} 1 = 2^N \tag{9.6}$$

   possible microstates for the lattice. Due to the *Principle of Indifference*, each microstate is equally likely and the probability for a distinguishable microstate r is therefore

$$p(\{s_i\}) = p(r) = \Omega^{-1} = 2^{-N} \qquad \text{for all possible microstates } r.$$

   We shorten the notation for the many sums over spins as follows:

$$\sum_{s_1=\pm1} \sum_{s_2=\pm1} \cdots \sum_{s_N=\pm1} \equiv \sum_{\{s_i\}} \equiv \sum_{r=1}^{\Omega} \tag{9.7}$$

   Note that the total energy E_r of the different microstates is by no means constant; for even L it can range from a minimum of $-2N$ to a maximum of $+2N$.

2. **No heat bath, fixed energy $E_r = E$ (microcanonical ensemble).** In this case, the question being asked no longer relates to the specific microstate r in which a system finds itself, but rather to its *total energy* E_r. Many different microstates r can have the same energy, $E_r = E$. In the microcanonical ensemble, the variable is therefore E, and all those microstates for which E_r equals the specified E are counted to determine $\Omega(E)$:

$$\Omega(E) = \sum_r \delta(E, E_r) \;=\; \sum_{\{s_i\}} \delta(E + J \textstyle\sum_{n.n.i,j} s_i s_j)\,, \tag{9.8}$$

where $\delta(E, E_r)$ is the Kronecker delta (remember the energies are discrete integer multiples of $J!$), which equals 1 when $E_r = E$ and 0 otherwise.

We can understand the microcanonical ensemble in terms of distinguishability as follows. Distinguishable microstates r, corresponding to the above Simple Sampling case, are binned by "energy states" characterised only by the energy E, i.e. microstates with the same energy are considered indistinguishable in the microcanonical ensemble case. Then $\Omega(E)$ is the *number of r-states which have the same* energy E and the corresponding probability for a system to have energy E is

$$p(E) = \frac{\Omega(E)}{\Omega} = \frac{\Omega(E)}{\sum_E \Omega(E)}. \tag{9.9}$$

The change from the set of N distinguishable microstates $r = \{s_1, \ldots, s_N\}$ to the single energy E is therefore a projection from an N-variable probability distribution onto a one-variable probability with total energy

$$p(r) = p(s_1, s_2, \ldots, s_N) \;\rightarrow\; p(E)\,,$$

in the same way that we projected N distinguishable coin tosses $X_1, \ldots, X_N$ with 2^N possible outcomes onto occupation numbers $n = 0, 1, \ldots, N$ with $N{+}1$ possible outcomes in the case of the binomial probability distribution.

The partition function $\Omega(E)$ can therefore be seen as an occupation number n_b, while the total number of states $\Omega = \sum_E \Omega(E)$ is the equivalent of N in ordinary binning and $p(E)$ of Eq. (9.9) corresponds to an occupation number probability. As before, $p(E)$ is not an experimental relative frequency but a real probability based on projection from the Principle of Indifference probability (9.7).

While the use of the fixed-energy case seems the obvious choice for simulation, this is not the case: it is very hard to implement systems with constant energy on the computer.

3. **Heat bath with a given (inverse) temperature** $\beta = 1/kT$: **canonical ensemble**. In this case, we do not keep the individual microstate energy constant but only the **average** energy, which is equivalent to keeping the temperature constant. The probability for **distinguishable microstates** r constrained by fixed β works out to be the Boltzmann distribution,

$$p(s_1, s_2, \ldots, s_N \,|\, \beta) = p(r \,|\, \beta) = \frac{1}{Z} e^{-\beta E_r} \tag{9.10}$$

where the partition function of all microstates r is

$$Z(\beta) = \sum_r e^{-\beta E_r} = \sum_{\{s_i\}} \exp\left[\beta J \sum_{n.n.i,j} s_i s_j\right]. \tag{9.11}$$

The ratio in (9.10) can be seen as the ratio of the unnormalised function $h(x) = e^{-\beta E_r}$ divided by its integral (or the sum, in the present case) $Z = \sum_r e^{-\beta E_r}$; see Eq. (**??**) above.

4. For the **indistinguishable** case and given constant temperature, we again bin all those microstates r which have the same total energy $E_r = E$; as above, there are $\Omega(E)$ such states for given E,[2]

$$p(E \,|\, \beta) = \frac{1}{Z} \Omega(E)\, e^{-\beta E} = \frac{\Omega(E)\, e^{-\beta E}}{\sum_E \Omega(E)\, e^{-\beta E}}, \tag{9.12}$$

[2]The exact way to derive this is as follows:

$$p(E \,|\, \beta) = \sum_r p(E_r \,|\, \beta)\, \delta(E, E_r) \;=\; Z^{-1} \sum_r e^{-\beta E_r} \delta(E, E_r) \;=\; Z^{-1} E^{-\beta E} \sum_r \delta(E, E_r) \;=\; Z^{-1} E^{-\beta E} \Omega(E)$$

where the partition function is written in terms of the occupation number $\Omega(E)$,

$$Z(\beta) = \sum_E \Omega(E) e^{-\beta E}. \tag{9.13}$$

$p(E \,|\, \beta)$ is the probability that a system has an energy E independent of which particular microstate was involved, but subject to the condition of fixed β. Note that $\Omega(E)$ is exactly the same partition function occurring in the microcanonical ensemble (9.8).

9.4 Importance sampling and the Ising model

We first note that the number of microstates grows tremendously quickly:

$$\Omega = 2^N = \begin{cases} 2^4 &= 16 & \text{for } L = 2 \\ 2^{25} &= 3.36 \times 10^7 & \text{for } L = 5 \\ 2^{100} &= 1.27 \times 10^{30} & \text{for } L = 10 \end{cases}$$

For a computer calculating a hundred million microstates per second, it would take approximately 3×10^{14} years to enumerate all the microstates for even the $L = 10$ case. In simulations, lattices of side length $L = 100$ or even $L = 500$ are not the exception but the rule. The implication is that is it completely impossible to access all possible microstates: Sampling and the use of Monte Carlo is a necessity.

Secondly, it is easy to show for very small $L \times L$ lattices that the energies of the possible microstates are *symmetric around zero.* (We almost always assume that there is no external magnetic field, $B = 0$.) This can easily be verified for a 2×2 lattice.

Furthermore, the distribution of energies can tend with increasing N to become more and more Gaussian, given that the energy is a sum of many small interactions. The Central Limit Theorem is not applicable at all temperatures because the individual spin states are not generally independent due to the interaction energy $s_i s_j$. Nevertheless, at temperatures away from the phase transition, the effect of the interaction is localised, so that summation nevertheless does result in something close to a gaussian.[3]

We can therefore make the statement that, for temperatures away from T_c, the number of microstates with the same total energy E has the functional form

$$\Omega(E) = c\, e^{-E^2/2\sigma^2} \tag{9.14}$$

because $\mu = E[E]$, the expectation value of the total energy, is equal to zero and where $\sigma^2 = E[E^2] - E[E]^2$ is the variance of the total energy. The constant $c = 2^N/\sigma\sqrt{2\pi}$ is found from $\int dE\, \Omega(E) = 2^N$, but its value is unimportant to us.

The implication of this distribution is that *simple sampling* in which we generate uniformly distributed microstates r *is a very ineffective strategy to generate microstates:*

- For the case of the *microcanonical ensemble*, the delta function in Eq. (9.8) implies that only a very small number of states r are accepted, namely those whose energy E_r happens to coincide with the given E, while the overwhelming majority of states will have an energy which is not equal to E and are therefore discarded.

[3] At the phase transition itself, the nearest neighbour interactions coalesce into long-range correlations, meaning that the individual spins are far from independent. This is reflected in the histograms both of $\Omega(E)$ and of M at the phase transition.

- For the *canonical ensemble.* the use of simple sampling is equally ineffective. In this case, we are looking for a sample of microstates r which have energies distributed according to (9.12). Eq (9.12) is, however, a gaussian with a peak at the point $E = -\beta\sigma^2$:

$$\begin{aligned} p(E \mid \beta) = \Omega(E)\, e^{-\beta E}/Z &= c\, e^{-\beta E - E^2/2\sigma^2}/Z \\ &= c\, e^{\sigma^2\beta^2/2}\ e^{-(E+\beta\sigma^2)^2/2\sigma^2}/Z. \end{aligned} \tag{9.15}$$

  The shift of the peak from zero to $-\beta\sigma^2$ is clearly dependent on the temperature $T = 1/\beta$; the smaller the temperature (i.e. the larger β), the further the gaussian $\Omega(E)\, e^{-\beta E}$ will deviate from the *simple sampling* gaussian around zero.

  This would in turn mean that it is highly ineffective to generate states by simple sampling, because the energies of the latter are distributed according to a gaussian centered around zero. Only a small minority of simple sampling states (those at low E) would thus make a contribution to the calculation.

It is clearly much more sensible to generate *a priori* a sample of microstates which is distributed according to $e^{-\beta E_r}$. *We are therefore looking for an algorithm which generates states r according to a distribution $f(r)$ approximately of the shape $e^{-\beta E_r}$ in such a way that the Boltzmann probability $p(r) = e^{-\beta E_r}/Z$ is precisely cancelled out.* According to the discussion of importance sampling in Section 8.5.4, averages $\langle h \rangle$ in a sample distributed according to $p(r)$ would simply be a sum over all the states r,[4]

$$\langle h \rangle_{IS} = \frac{1}{n} \sum_{r=1}^{n} h(r) \frac{p(r)}{f(r)} \qquad r \sim f(r). \tag{9.16}$$

In particular, Eqs. (8.41) and (8.48) tell us that, if it is possible to construct $f(r)$ in such a way that it approaches $p(r)$, the ratio $p(r)/f(r)$ will tend to 1 and the sample average of h will become equal to $(1/n)\sum_{r=1}^{n} h(r)$ while avoiding all the problems of simple sampling.

9.5 Markov Chain Monte Carlo

9.5.1 Introducing a time variable

Importance sampling can in principle handle the problem of the shifting of the peak of the energy distribution of the states; the question is, however, how that is achieved in practice: how do we construct $f(r)$? As we have seen, the number of variables $s_1, \ldots, s_N$ grows with the square of the lattice size, and the number of states 2^N is huge and the corresponding probability 2^{-N} is tiny. Dividing numbers of this magnitude can easily make importance sampling a lost cause: how does one dream up a computable probability $f(r) = f(s_1, \ldots, s_N)$ which matches the largely unknown $p(r)$ we want to sample? We often do not even know in which parts of the huge outcome space $\mathcal{A}(s_1, \ldots, s_N)$ the peak of $p(r)$ is to be found. For a discussion of this problem, see MacKay Chapter 29.

The solution of Metropolis et al. was to introduce a time coordinate t and a corresponding *stochastic process* i.e. probabilities of a set of variables depending on a fictional time t. This "time" may correspond to what we normally call physical time, and in Bayesian terms we may identify is as a "information time", but it is really only an auxiliary variable designed to sample $f(r)$ more efficiently. In the simplest case, t is a positive integer $0, 1, \ldots$; it can be considered the index of a `for`-loop. The essence of stochastic processes is that we consider the probability $f(r \mid \beta)$ to be a function of t; indeed for every t we have a different variable r_t and a different probability

$$f(r_t \mid \beta) \qquad t = 0, 1, 2, \ldots$$

[4] We write n for the Monte Carlo sample size while N is the number of lattice sites.

with $f(r_{t=\infty} \,|\, \beta) = p(r \,|\, \beta)$ the desired importance sampling reference probability. In other words, the sequence of probabilities $f(r_t \,|\, \beta)$ is to be designed in such a way that it converges to the right one, where "right one" for the Ising model would be the Boltzmann distribution,

$$\lim_{t\to\infty} f(r_t \,|\, \beta) = p(r \,|\, \beta) = \frac{e^{-\beta E_r}}{Z}. \tag{9.17}$$

To repeat: $f(r_t \,|\, \beta)$ is equivalent to $f(x)$ in importance sampling, and the idea is for f to be designed in such a way that it approaches the desired $p(x) = p(r \,|\, \beta)$ which is the Boltzmann distribution for a given inverse temperature β.

Let us drop the β in the conditional for the moment and simply write $f(r_t)$. Then the rules of marginalisation $f(x) = \sum_y f(x, y)$ imply that we can always expand $f(r_{t+1})$ in terms of the joint probability,

$$f(r_{t+1}) = \sum_{r_t} f(r_{t+1}, r_t) = \sum_{r_t} f(r_{t+1} \,|\, r_t)\, f(r_t) \tag{9.18}$$

We can write this in terms of vectors and a matrix by arranging the probabilities of all possible outcomes $f(r_t{=}r), r = 1, 2, \ldots \Omega$ in a very long column vector which we can call $\boldsymbol{F}_t$. Similarly we can arrange the probabilities for all outcomes for r_{t+1} as the column vector $\boldsymbol{F}_{t+1}$. The conditional probability $f(r_{t+1}{=}k \,|\, r_t{=}\ell)$, written as matrix $\mathbb{W}$ with components $W_{k\ell}$ then connects any component $(\boldsymbol{F}_t)_\ell = f(r_t{=}\ell)$ at time t to any component $(\boldsymbol{F}_{t+1})_k = f(r_{t+1}{=}k)$ at time $t{+}1$, and the above identity becomes

$$\boldsymbol{F}_{t+1} = \mathbb{W}\boldsymbol{F}_t, \tag{9.19}$$

where the matrix $\mathbb{W}$ has components

$$W_{k\ell} = f(r_{t+1}{=}k \,|\, r_t{=}\ell) \tag{9.20}$$

with k, ℓ shorthand for discrete states in $\mathcal{A}(r_t) = \mathcal{A}(r_{t+1})$, the outcome space of the process which is assumed discrete and which is time-stationary, i.e. independent of t. The change of the probability for the system to be in microstate k is the difference

$$\frac{\Delta f(r_t)}{\Delta t}\Big|_k = f(r_{t+1}{=}k) - f(r_t{=}k) \tag{9.21}$$

which we write as $d\boldsymbol{F}/dt$ only for convenience; Δt is not infinitesimal. In vector form, this reads

$$\frac{d\boldsymbol{F}}{dt} = \boldsymbol{F}_{t+1} - \boldsymbol{F}_t. \tag{9.22}$$

If $\mathbb{I}$ is the identity matrix, we obtain after inserting (9.19),

$$\frac{d\boldsymbol{F}}{dt} = (\mathbb{W} - \mathbb{I})\boldsymbol{F} \tag{9.23}$$

so that the stochastic process can be fully expressed in matrix-vector notation.

9.5.2 Equilibrium and detailed balance

So-called *equilibrium* is reached at some time t_{eq} if and only if the probabilities no longer change,

$$0 = \frac{d\boldsymbol{F}_t}{dt} \qquad \forall\, t > t_{\text{eq}} \tag{9.24}$$

or straightforwardly, in matrix and component form,

$$\boldsymbol{F}_{t+1} = \boldsymbol{F}_t \tag{9.25}$$

$$f(r_{t+1}{=}k) = f(r_t{=}k) \tag{9.26}$$

The latter can of course be written in terms of marginals,

$$\sum_{r_t} f(r_{t+1}{=}k, r_t) = \sum_{r_{t+1}} f(r_{t+1}, r_t{=}k) \qquad \forall\, k,$$

which we may as well write as sums over the outcomes ℓ,

$$\sum_{\ell} f(r_{t+1}{=}k, r_t{=}\ell) = \sum_{\ell} f(r_{t+1}{=}\ell, r_t{=}k) \tag{9.27}$$

or, using the product rule,

$$\sum_{\ell} f(r_{t+1}{=}k \,|\, r_t{=}\ell)\, f(r_t{=}\ell) = \sum_{\ell} f(r_{t+1}{=}\ell \,|\, r_t{=}k)\, f(r_t{=}k). \tag{9.28}$$

On the other hand, we have from Eqs. (9.19) and (9.25) that at equilibrium

$$\boldsymbol{F}_t = \mathbb{W}\boldsymbol{F}_t, \tag{9.29}$$

which resembles the general evolution $\boldsymbol{F}_{t+1} = \mathbb{W}\boldsymbol{F}_t$ of (9.19) but is clearly not the same. Eq. (9.29) implies that in equilibrium we have an eigensystem $\mathbb{W}\boldsymbol{F}_t = \boldsymbol{F}_t$, and the equilibrium probability $\boldsymbol{F}_{\text{eq}} = \boldsymbol{p}$ will therefore be the eigenvector of $\mathbb{W}$ for the eigenvalue $\lambda = 1$. This may seem sufficient to fully determine the problem: Given some choice of transition probability $f(r_{t+1} \,|\, r_t)$ and hence matrix $\mathbb{W}$, we find its eigenvector with eigenvalue $\lambda = 1$. However, there is a complication: it is possible that the system falls into an "limit cycle" state in which a state evolving over c steps

$$\boldsymbol{F}_{t+c} = \mathbb{W}^c\boldsymbol{F}_t, \tag{9.30}$$

returns to exactly the same initial state,

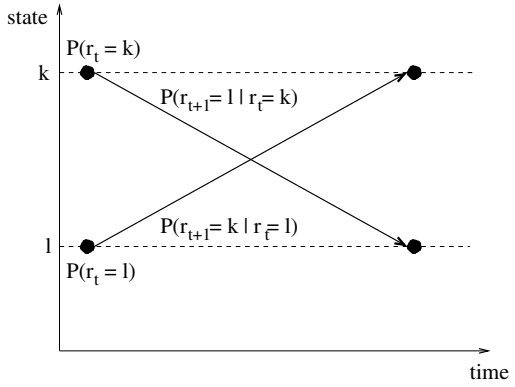
$$\boldsymbol{F}_{t+c} = \boldsymbol{F}_t = \mathbb{W}^c\boldsymbol{F}_t. \tag{9.31}$$

This situation is undesirable since a limit cycle means that there is no single equilibrium state and no equilibrium. To ensure that we get real equilibrium rather than such cycles or oscillations, we must impose an additional condition termed **detailed balance**. In component form, detailed balance demands that the variables r_t and r_{t+1} must be exchangeable,

$$f(r_{t+1}{=}k, r_t{=}\ell) \stackrel{!}{=} f(r_{t+1}{=}\ell, r_t{=}k) \qquad \forall\, k, \ell \tag{9.32}$$

$$\text{or} \quad f(r_{t+1}{=}k \,|\, r_t{=}\ell)\, f(r_t{=}\ell) \stackrel{!}{=} f(r_{t+1}{=}\ell \,|\, r_t{=}k)\, f(r_t{=}k) \tag{9.33}$$

which closely resemble the equilibrium equations (9.27)–(9.28) except that the equilibrium equations involve *sum over* ℓ while detailed balance imposes equal terms for *every single* ℓ. The condition of detailed balance can be represented as the graph below, where the dots represent probabilities $f(r)$ and the arrows represent transition probabilities $W(r_{t+1} \,|\, r_t)$.



Eq. (9.33) comes close to the definition of equilibrium used in chemistry, namely that there is no net change in the concentration (probability) of particular chemicals A and B if and when the concentration $p(B)$ times the rate of change from B to A is equal to the concentration $p(A)$ time the rate of change from A to B.

Note that equilibrium does not imply that the conditional probabilities to change from any state k to any other state ℓ is zero. It merely states that the total changes back and forth cancel out. Similarly, equilibrium does not imply that $\mathbb{W} = \mathbb{I}$ or, in ordinary language, that $f(r_{t+1}{=}k \,|\, r_t{=}\ell) = 0$ for all $k \neq \ell$. The solution $\mathbb{W} = \mathbb{I}$ exists but that particular solution would mean that no microstate ever changes into another.

We are now ready to motivate the Metropolis algorithm. Rearranging the factors in (9.33),

$$\frac{f(r_{t+1}{=}\ell \,|\, r_t{=}k)}{f(r_{t+1}{=}k \,|\, r_t{=}\ell)} = \frac{f(r_t{=}\ell)}{f(r_t{=}k)} \tag{9.34}$$

we see that the ratio of transition probabilities equals the ratio $f(r_t{=}\ell)/f(r_t{=}k)$. If we now simply impose our goal, namely that $f(r_t{=}\ell)/f(r_t{=}k)$ must be equal to the ratio of desired equilibrium probabilities

$$\frac{f(r_t{=}\ell)}{f(r_t{=}k)} \stackrel{!}{=} \frac{p(r_t{=}\ell)}{p(r_t{=}k)}, \tag{9.35}$$

after some finite t, then by design

$$\frac{f(r_{t+1}{=}\ell \,|\, r_t{=}k)}{f(r_{t+1}{=}k \,|\, r_t{=}\ell)} \stackrel{!}{=} \frac{p(r_t{=}\ell)}{p(r_t{=}k)}. \tag{9.36}$$

9.5.3 Application to Ising model

The above is quite general and applied to many different problems. Before dealing with the general case, we consider the specific application to the Ising model below, for which the desired distribution $p(r)$ is the Boltzmann equilibrium distribution $p(r \,|\, \beta) = e^{-\beta E_r}/Z$, we would impose the condition

$$\frac{f(r_{t+1}{=}\ell \,|\, r_t{=}k)}{f(r_{t+1}{=}k \,|\, r_t{=}\ell)} = \frac{e^{-\beta E_\ell}}{e^{-\beta E_k}}. \tag{9.37}$$

Based on this insight, Metropolis et al. came up with the following algorithm. The version below is couched in the language of the Ising model but is actually applicable to a wide range of similar problems.

Metropolis algorithm for Ising Model:
Given an Ising lattice which is in a particular state $r_t = \{s_1, \ldots, s_i, \ldots, s_N\}$ *with energy* $E_r \equiv E_{r_t}$, implement the following loop algorithm:

1. Randomly flip a single spin i from s_i to $s'_i = -s_i$ to generate a candidate state $r' = \{s'_1, \ldots, s'_i, \ldots, s'_N\} = \{s_1, \ldots, -s_i, \ldots, s_N\}$ which differs from state r only in that one spin orientation, $s'_i = -s_i$.

2. Calculate the difference between the candiate state's energy $E_{r'}$ and the old energy E_r. Since the energy difference depends only on the change in interaction between spin i and its four nearest neighbours, $E_{r'} - E_r$ of the entire lattice can

be determined by considering just these four interaction terms,

$$\begin{aligned}\Delta E &= E_{r'} - E_r = -J\sum_{j|i} s_i' s_j + J\sum_{j|i} s_i s_j = J(-s_i' + s_i)\sum_{j|i} s_j \\ &= 2Js_i \sum_{j|i} s_j\,, \end{aligned} \tag{9.38}$$

where the j-sum runs over only the four nearest neighbours of i.

3. If $\Delta E \leq 0$, accept the new spin orientation of spin i and admit the candidate state r' as the new state r_{t+1} in your chain.

4. If $\Delta E > 0$, generate a uniform random number u between 0 and 1. If $u < e^{-\beta\,\Delta E}$, accept the flip and admit the new configuration r' as r_{t+1} into the chain. If $u \geq e^{-\beta\,\Delta E}$, accept the old state $r_t \equiv r$ as new member r_{t+1} of the chain.

5. Return to 1.

To clarify this algorithm, we use our standard rules of probability. The probability of the new state r_{t+1} depends both on the old state r_t and the candidate state r', so with marginalisation and the product rule we have

$$f(r_{t+1}\,|\,r_t) = \sum_{r'} f(r_{t+1}, r'\,|\,r_t) = \sum_{r'} f(r_{t+1}\,|\,r', r_t)\, p(r'\,|\,r_t) \tag{9.39}$$

The Metropolis flip-one-spin prescription implies that r' differs from r_t by the change of only one spin at one lattice point i, so there are N possible candidate states. Let the selection probability of that lattice point be uniform, so that[5]

$$f(r'\,|\,r_t) = \frac{1}{N} \qquad \forall \text{ lattice points } i, \tag{9.40}$$

while the remainder of the prescription is captured in terms of the energy difference between the current state r_t and the candidate state r',

$$\Delta E_t = E_{r'} - E_{r_t}, \tag{9.41}$$

as

$$f(r_{t+1}{=}r'\,|\,r', r_t) = \theta[-\Delta E_t] + \theta[\Delta E_t]\, e^{-\beta\Delta E_t} = \min[1, e^{-\beta\,\Delta E_t}] \tag{9.42}$$

with the Heaviside theta functions enforcing the conditions that r' is accepted with probability 1 if $\Delta E_t \leq 0$ but with probability $e^{-\beta\Delta E_t}$ when $\Delta E_t > 0$. With the help of the identity $1 = \theta[x] + \theta[-x]$, the keep-old-state probability is then

$$\begin{aligned} f(r_{t+1}{=}r_t\,|\,r', r_t) &= 1 - f(r_{t+1}{=}r'\,|\,r', r_t) \\ &= \theta[\Delta E_t]\left(1 - e^{-\beta\Delta E_t}\right). \end{aligned} \tag{9.43}$$

which is 0 when $\Delta E_t < 0$ as required by the Metropolis algorithm.

[5] We can write this in exact but cumbersome notation as

$$p(s_1', s_2', \ldots, s_i', \ldots s_N'\,|\,s_1, s_2, \ldots, s_i, \ldots, s_N) = (1/N)\delta(s_i', -s_i)\prod_{j\neq i}\delta(s_j', s_i)$$

It is easy to show that this algorithm implies that, whenever equilibrium is reached and detailed balance applies, $f(r_t)$ does indeed asymptotically approach the Boltzmann distribution $p(r\,|\,\beta)$. Compare two microstates ℓ and k which are related by a single spin flip, and assume that E_ℓ is larger than E_k. For the transition $\ell \to k$, $(E_k - E_\ell) = \Delta E < 0$, and the new state k is therefore accepted with probability $1/N$ which includes $p(r'\,|\,r_t)$. For the opposite transition $k \to \ell$, $(E_\ell - E_k) = \Delta E > 0$, in which case the new state ℓ is accepted with a probability $f(r_{t+1}{=}r'{=}\ell\,|\,r_t{=}k) = (1/N)\,e^{-\beta(E_\ell - E_k)}$. The ratio is therefore

$$\frac{f(r_{t+1}{=}r'{=}\ell\,|\,r_t{=}k)}{f(r_{t+1}{=}r'{=}k\,|\,r_t{=}\ell)} = \frac{(1/N)\,e^{-\beta(E_\ell - E_k)}}{(1/N)} = \frac{e^{-\beta E_\ell}}{e^{-\beta E_k}} \quad = \frac{p(\ell\,|\,\beta)}{p(k\,|\,\beta)} \tag{9.44}$$

showing that $f(r_t)$ does follow the Boltzmann distribution in equilibrium.[6] **By means of a pseudo time series, we can therefore generate a set of microstates r_t which are distributed according to the Boltzmann distribution of the given β.**

The intuitive explanation of the algorithm would be to say that a transition to a lower energy is always permitted, while a transition to a higher energy is discouraged but not prohibited. The latter is important to prevent the system getting stuck in a local minimum of the energy which is still higher than the true equilibrium energy. Possible evolution to a state with higher energy is therefore a necessity.

9.5.4 General MCMC and detailed balance

As before, we aim to find expectation values of some scalar function $h(x)$ with respect to a known probability $p(x)$ which, however, cannot be sampled directly. As already argued, we find a numerical estimate for $E_p(h(x))$ by introducing a Markov process $f(X_t\,|\,X_{t-1})$ with a discrete time index $t = 0, 1, 2, \ldots$ and outcomes $x_t \in \mathcal{A}(X)$, where $\mathcal{A}(X)$ is the same for all X_t. The derivations work equally in higher-dimensional spaces but the mathematics for one-dimensional outcome spaces is sufficient. We note in passing that there are prerequisites such as ergodicity; in a nutshell, the Markov process must be capable of reaching all states in $\mathcal{A}(X)$ within some finite time. The detailed balance requirement is

$$\frac{f(X_{t+1}{=}\ell\,|\,X_t{=}k)}{f(X_{t+1}{=}k\,|\,X_t{=}\ell)} \stackrel{!}{=} \frac{p(X_t{=}\ell)}{p(X_t{=}k)} \qquad \forall\, k, \ell \in \mathcal{A}(X),$$

where k, ℓ are shorthand for specific states x_t in the sampling space. We have already replaced the equilibrated distribution $f(x)$ by the desired target distribution $p(x)$ as before. To simplify the notation, we write the new-state variable as $Y = X_{t+1}$, the current-state variable as $X = X_t$.

Let $\overline{\delta}_{k\ell}$ be the anti-Kronecker delta which is 1 whenever $k \neq \ell$ and 0 otherwise. Then for any k, ℓ, the sum of Kronecker and anti-Kronecker is always $\delta_{kl} + \overline{\delta}_{k\ell} = 1$ and the transition probability can be split into

$$f(X_{t+1}{=}\ell\,|\,X_t{=}k) \;=\; f(Y{=}\ell\,|\,X{=}k) = \delta_{k\ell}\, f(Y{=}k\,|\,X{=}k) + \overline{\delta}_{k\ell}\, f(Y{=}\ell\,|\,X{=}k). \tag{9.45}$$

We now consider the details of the transition from X_t to X_{t+1}. At time t, we already have a sequence of outcomes or "samples" $x_1, x_2, \ldots, x_t$ drawn from $X_1, X_2, \ldots, X_t$, for which k and ℓ are specific examples. To find the next sample x_{t+1} we again introduce a candidate-state variable X'. With some probability $f(X_{t+1}\,|\,X', X_t)$ as described below, the candidate state will be *accepted*, meaning that $x_{t+1} = x'$ or *rejected*, which implies $x_{t+1} = x_t$. In order to introduce the candidate variable, we marginalise and use the product rule for the accept part,

$$\overline{\delta}_{k\ell}\, f(Y{=}\ell\,|\,X{=}k) = \overline{\delta}_{k\ell} \sum_j f(Y{=}\ell\,|\,X'{=}j, X{=}k)\, f(X'{=}j\,|\,X{=}k) \tag{9.46}$$

[6] The case where E_ℓ is equal to E_k implies $e^{-\beta\Delta E} = 1$ so that it does not matter how we handle it.

where j is yet another specific outcome in $\mathcal{A}(X)$. This we rewrite in less transparent but more conventional notation in terms of the *proposal probability* $q(X' \,|\, X)$ and the *acceptance probability* $\alpha(Y \,|\, X', X)$ so that Eqs. (9.45)–(9.46) become[7]

$$f(Y{=}\ell \,|\, X{=}k) = \delta_{k\ell}\, f(Y{=}k \,|\, X{=}k) + \overline{\delta}_{k\ell} \sum_j \alpha(Y{=}\ell \,|\, X'{=}j, X{=}k)\, q(X'{=}j \,|\, X{=}k). \tag{9.47}$$

The marginalisation sum over j has only one nonzero term, because a state can be accepted only if that state is proposed, so if $Y{=}\ell$, then only the term $X'{=}j{=}\ell$ could lead to acceptance. Hence

$$f(Y{=}\ell \,|\, X{=}k) = \delta_{k\ell}\, f(Y{=}k \,|\, X{=}k) + \overline{\delta}_{k\ell}\, \alpha(Y{=}\ell \,|\, X'{=}\ell, X{=}k)\, q(X'{=}\ell \,|\, X{=}k). \tag{9.48}$$

We now introduce detailed balance as a requirement, which in the simplified notation reads

$$\frac{f(Y{=}\ell \,|\, X{=}k)}{f(Y{=}k \,|\, X{=}\ell)} \stackrel{!}{=} \frac{p(X{=}\ell)}{p(X{=}k)} \qquad \forall\, k, \ell \in \mathcal{A}(X). \tag{9.49}$$

It is trivially satisfied for the case $k = \ell$ (or $\delta_{k\ell}$ in the above equation), so we concentrate on the second part $\overline{\delta}_{k\ell}$ of unequal outcomes $k \neq \ell$. For this case, Eq. (9.49) becomes

$$\frac{\alpha(Y{=}\ell \,|\, X'{=}\ell, X{=}k)\; q(X'{=}\ell \,|\, X{=}k)}{\alpha(Y{=}k \,|\, X'{=}k, X{=}\ell)\; q(X'{=}k \,|\, X{=}\ell)} \stackrel{!}{=} \frac{p(X{=}\ell)}{p(X{=}k)} \qquad \forall\, k \neq \ell. \tag{9.50}$$

Under detailed balance, the *acceptance ratio* of the two acceptance probabilities is hence

$$R(k, \ell) = \frac{\alpha(Y{=}\ell \,|\, X'{=}\ell, X{=}k)}{\alpha(Y{=}k \,|\, X'{=}k, X{=}\ell)} = \frac{q(X'{=}k \,|\, X{=}\ell)\; p(X{=}\ell)}{q(X'{=}\ell \,|\, X{=}k)\; p(X{=}k)}. \tag{9.51}$$

Note carefully the order in which specific outcomes k and ℓ appear in the different ratios.

Rejection probability: Eq. (9.48) immediately gives us a way to calculate the rejection probability. Summing over ℓ on both sides, we have $1 = f(Y{=}k \,|\, X{=}k) - \sum_\ell \overline{\delta}_{k\ell}\, \alpha(Y{=}\ell \,|\, X'{=}\ell, X{=}k)\, q(X'{=}\ell \,|\, X{=}k)$ so that the rejection probability can be calculated from

$$f(Y{=}k \,|\, X{=}k) = 1 - \sum_\ell \overline{\delta}_{k\ell}\, \alpha(Y{=}\ell \,|\, X'{=}\ell, X{=}k)\, q(X'{=}\ell \,|\, X{=}k). \tag{9.52}$$

9.5.5 Metropolis-Hastings algorithm

The derivation up to this point holds for any proposal probability q and any acceptance probability α. The specific choice made within the Metropolis-Hastings algorithm for the acceptance probability is

$$\alpha(Y{=}\ell \,|\, X'{=}\ell, X{=}k, \mathcal{H}_{\text{MH}}) = \min\left[1\,,\; \frac{q(X'{=}k \,|\, X{=}\ell)\; p(X{=}\ell)}{q(X'{=}\ell \,|\, X{=}k)\; p(X{=}k)}\right] \tag{9.53}$$

so that the acceptance ratio is

$$R(k, \ell, \mathcal{H}_{\text{MH}}) = \frac{\alpha(Y{=}\ell \,|\, X'{=}\ell, X{=}k, \mathcal{H}_{\text{MH}})}{\alpha(Y{=}k \,|\, X'{=}k, X{=}\ell, \mathcal{H}_{\text{MH}})} = \frac{\min\left[1\,,\; \dfrac{q(X'{=}k \,|\, X{=}\ell)\; p(X{=}\ell)}{q(X'{=}\ell \,|\, X{=}k)\; p(X{=}k)}\right]}{\min\left[1\,,\; \dfrac{q(X'{=}\ell \,|\, X{=}k)\; p(X{=}k)}{q(X'{=}k \,|\, X{=}\ell)\; p(X{=}\ell)}\right]}. \tag{9.54}$$

This seemingly complicated object is easily parsed by considering the two cases of $q(X'{=}k \,|\, X{=}\ell)\; p(X{=}\ell)$ being larger or smaller than $q(X'{=}\ell \,|\, X{=}k)\; p(X{=}k)$. In both cases, Eq. (9.54) reduces to the detailed balance equation (9.51).

[7] In the literature, the acceptance probability notation $\alpha(Y \,|\, X)$ normally omits the X' since Y and X must always have the same outcome, but we prefer to keep X' for clarity.

Metropolis-Hastings algorithm

Physics input: The target probability $p(x)$
User-set parameters: x_0, T_{eq}, choice of proposal distribution $q(x_{t+1} \,|\, x_t)$
Burn-in phase: Run MH for-loop until equilibrium has been attained
MH for-loop:
for $t = 1, \ldots, T_{\text{eq}}$ **do**
    Generate candidate sample $x' \sim q(x' \,|\, x_t)$
    Calculate the acceptance ratio and acceptance probability

$$R(x', x_t) = \frac{q(x_t \,|\, x')}{q(x' \,|\, x_t)} \cdot \frac{p(x')}{p(x_t)} \qquad \alpha(x_{t+1} \,|\, x', x_t) = \min[1, R(x', x_t)]$$

    Assign the next state x_{t+1} as

$$x_{t+1} = \begin{cases} x' & \text{i.e. accept with probability } \alpha \\ x_t & \text{i.e. reject with probability } 1 - \alpha \end{cases}$$

    In detail, the assignment proceeds as follows:
    if $R > 1$ **then**
        $x_{t+1} = x'$
    else
        Generate a uniformly-distributed random number u between 0 and 1.
        Assign the next state x_{t+1} as

$$x_{t+1} = \begin{cases} x' & \text{if } u < \alpha(x_{t+1} \,|\, x', x_t), \text{ or} \\ x_t & \text{otherwise.} \end{cases}$$

    end if
end for
Calculate Autocorrelation function and autocorrelation time τ
Run above for-loop in equilibrium, using only every (2τ)-th sample
Return Sample of independent points distributed according to $p(x)$

9.5.6 Properties and disadvantages of MCMC

We conclude with a list of relevant issues and remarks.

1. **The goal: a set of independent samples distributed according to** $p(x)$. Before launching into the details of properties, advantages and disadvantages of MCMC, we remind ourselves of the overall goal. We are aiming to obtain a set of *simdata*

$$\mathcal{D} = \{x_i\}_{i=1}^{N} \tag{9.55}$$

   which must have two crucial properties:

   i. The x_i must be *distributed according to the desired target probability* $p(x)$.
   ii. The simdata should be *independent* in the sense that any numerical estimate (of the probability $p(x)$ itself or any derived quantity such as moments and expectation values) should converge to the corresponding iid quantity. If, for example the sample mean of the product does not converge to the product of the first moments $\lim_{N\to\infty} \langle x_i x_j \rangle \neq E_p(x)^2$, the simdata could not be termed independent.

The statistical description of this sample by for example sample means, sample variances etc. which truthfully reflect the corresponding properties of $p(x)$ are the final goal of the entire effort.

2. **MCMC disadvantages** Neither of these goals will be automatically fulfilled without application of special measures and great care. While the MCMC family of algorithms represent a fundamental and elegant solution to the problem of sampling $p(x)$, it does create complications with which the user must deal.

   i. **Dependence of states:** For the purposes of calculating credible sample averages, the underlying states have to be independent. Two states x_t and x_{t+1} are clearly far from independent, because the Markov probability for x_{t+1} is based on knowledge of x_t. The only way to even approach independence would be multiple transitions from x_t to $x_{t+\delta}$ with δ large enough to obscure the implicit dependence.

   There is no simple way to determine δ except by simulation. One of a number of numerical methods is the "autocorrelation time" as set out in more detail below.

   ii. **Burn-in:** By experience and from theory, it is highly improbable that the initial states are already distributed according to $p(x)$, because the choice of x_0 may not be typical for it at all. The evolution from x_0 to samples which do represent $p(x)$ faithfully is a necessary evil called burn-in. What exactly the adjective *faithfully* in the statement "the states faithfully represent p" is defined depends the specific problem at hand. Also, there is no unique prescription how to even measure or quantify whether a particular simdata actually does represent the target distribution. In low dimensions, comparisons of histograms with the actual plot of $p(x)$ may still work, but this becomes impossible in high-dimensional probabilities which commonly occur in MCMC work.

   iii. In the case of the Ising model, the accepted measure of completed burn-in is the achievement of a measure of invariance (constancy) of the magnetisation as a function of t, but even this criterion is fraught with uncertainties at temperatures near the phase transition.

   As a safeguard, one can start two separate stochastic processes with different seeds. True equilibrium is plausible only if the magnetisation of both processes are not only approximately constant in time, but also approximately equal to each other.

3. **The role of the acceptance probability**

   i. The listed disadvantages imply that MCMC codes will often require considerable CPU time, implying higher costs and project delays. Eliminating unnecessary inefficiencies is therefore an essential component of good coding.

   ii. Given that new parts of $\mathcal{A}(X)$ can only be explored if the MCMC particle efficiently reaches and explores those regions where $p(x)$ is comparatively large, a high acceptance probability is very desirable. A low acceptance ratio implies many rejections and hence a slow exploration of the sampling space $\mathcal{A}(X)$.

   iii. We therefore want to maximise the generic acceptance ratio without violating the constraint that individual probabilities may not exceed 1.

   iv. The selection probability $q(x' \,|\, x)$ should resemble $p(x)$ as closely as possible. The ideal choice would be one where q has such a form that samples are produced exactly proportionally to $p(x)$, so that acceptance is almost always close to maximal.

   The most obvious result of this consideration is to construct a $q(x' \,|\, x)\, p(x)$ ratio and then set the acceptance probability α equal to that ratio if it is smaller than 1 and set α to 1 if the ratio exceeds 1. This is exactly the insight which led to the Metropolis-Hastings algorithm and the use of the $\min[1, R(k, \ell)]$ prescription.

4. **Unnormalised probabilities:** The fact that detailed balance only requires ratios of probabilities is very helpful, since there is no need to know the normalisation constant of an unnormalised probability q because that normalisation constant cancels in the ratio. The ratio formulation only requires that q can be calculated in some way, without need to check its normalisation or find its normalisation constant.

5. We finally observe that a part of the flexibility of MCMC is the result that detailed balance is automatically satisfied for the reject-case, $k = \ell$. This means that the accept-probabilities have some freedom which can be used for optimisation, because such adjustments can be accommodated within the overall normalisation constraint by an equivalent adjustment in the magnitude of the reject-probability.

9.6 Time autocorrelation

It is not obvious how to determine whether two states in the spin-flip chain are independent of each other. As mentioned above, the two states before and after a single Monte Carlo sweep are still quite similar, and one must do δ such sweeps before you can talk of meaningful differences between successive lattice configurations (microstates) x_k and x_{k+1} in the collected sample.

The *time autocorrelation*, the covariance for a fixed time difference Δt, is a way to approach this issue quantitatively. The unit of time is here defined as one MC sweep, and once we have chosen a maximum number of sweeps $t_{\max}$ of the simulation (starting the count after the preliminary equilibration phase) we define the time autocorrelation as

$$\begin{aligned} C(\Delta t) &= \langle m(t')\, m(t'-\Delta t)\rangle - \langle m\rangle_1 \langle m\rangle_2 \\ &= \frac{1}{t_{\max}-\Delta t} \sum_{t'=\Delta t}^{t_{\max}-1} \big[m(t')\, m(t'-\Delta t)\big] \;-\; \langle m\rangle_1 \langle m\rangle_2 \qquad \Delta t = 0, 1, \ldots \end{aligned} \tag{9.56}$$

where we calculate the average magnetisation per spin $\langle m\rangle$ as

$$\langle m\rangle_1 = \frac{1}{t_{\max}-\Delta t} \sum_{t'=0}^{t_{\max}-\Delta t-1} m(t') \tag{9.57}$$

$$\langle m\rangle_2 = \frac{1}{t_{\max}-\Delta t} \sum_{t'=\Delta t}^{t_{\max}-1} m(t') \tag{9.58}$$

The limits on the sums ensure that $t' + \Delta t$ never becomes larger than $t_{\max}$.

Experience and theory tells us that the autocorrelation looks like an exponential,

$$C(\Delta t) = C(0)\, e^{-\Delta t/\tau}, \tag{9.59}$$

where the parameter τ represents the "correlation time" with $C(\Delta t{=}\tau) = C(0)e^{-1}$. The number of MC sweeps which need to pass before a new state is accepted as independent is normally taken as 2τ,

$$\delta = 2\tau. \tag{9.60}$$

The correlation time does vary for different temperatures: far from the phase transition, τ is small, but in the interesting region near the critical temperature T_c' the correlation time τ becomes quite large. It is obviously not necessary to calculate or plot $C(\Delta t)$ for $\Delta t \gg \tau$.

Important note on buffers:
The prescription (9.56) works well if the entire array of magnetisations is stored. However, when simulations become very long, this becomes impractical. In this case, a better solution is to only store $t_{\rm buffer}$ values of the magnetization, where $t_{\rm buffer}$ is the longest autocorrelation you want to calculate; normally $t_{\rm buffer} \ll t_{\rm max}$. In the case of using buffers with $t_{\rm buffer}$, (9.56) changes to

$$C(\Delta t) \;=\; \frac{1}{t_{\rm buffer}} \sum_{t'=\Delta t+1}^{t_{\rm buffer}} \big[m(t')\, m(t'-\Delta t)\big] \;-\; \langle m\rangle_1 \langle m\rangle_2 \qquad \Delta t = 0, 1, \ldots t_{\rm bmax} \tag{9.61}$$

while $\langle m\rangle_1$ and $\langle m\rangle_2$ etc are similarly modified. While this may seem like a minor detail, it turns out to be crucial for the correct implementation of autocorrelations. The maximum buffer length $t_{\rm buffer}^{\rm max}$ should normally be less than half of $t_{\rm max}$ and can often be much less than that.