Computer Intensive Statistics: APTS 2025–26 Preliminary Material

Andi Q. Wang (andi.wang@warwick.ac.uk)

June 2026

Contents

1 Introduction

The principal aim of these notes is to provide context and motivation for the APTS Computer Intensive Statistics Module, and to make the module as self-contained as is feasible. Statistics is a broad discipline and the APTS cohort naturally has a diverse range of backgrounds. If you have attended the earlier APTS modules this year, especially *Statistical Computing* and *Statistical Inference*, then you should be well-prepared for this module.

Although it is likely that everyone attending the module will know everything they need to follow the lectures, I highly recommend you familiarise yourself with some implementation and basic computer programming. For historical reasons, for the examples in this module we will be largely making use of the R programming language (R Core Team 2013), but you are free during the labs to use any language you prefer. If you haven't got any real experience of programming, please try the R Foundations course before the start of the APTS week itself.

Several appendices are provided. You might very well already know everything in these appendices—if that's the case, then great. If you have a less conventional statistical background and haven't managed to attend the earlier APTS modules then you may find some parts of these notes less familiar in which case some references are provided, but rest assured that the module should be accessible to anyone who is pursuing a PhD in any aspect of statistics or applied probability (both interpreted broadly). Appendix A provides a compact summary of some statistical tasks which we will aim to address in this module. It is likely that anyone pursuing a PhD in statistics—especially anyone who has attended an APTS module on *Statistical Inference*—will be familiar with this material, and (re)reading the notes provided for that module would be good preparation for the present one, but it is also convenient to have a compact summary on hand. Appendix B summarises some basic notions of convergence of stochastic quantities. This is here only to ensure that everyone has had the opportunity to see these ideas before the week. If everything it contains is obvious to you then great; if it's not then don't panic, we will make limited use of these results and their friends to motivate and justify some of what we do in this module. But we won't spend time proving them or focussing heavily on their technical consequences[1].

Appendix C, on the other hand, is likely to contain material which is entirely new to most of you. It contains an introduction to a recently proposed perspective on probability, based on category theory. During the course, use of categorical probability will streamline many proofs which would otherwise consist of tedious strings of integrals. Understanding those proofs will *not* require any prior knowledge of categorical probability or category theory. Appendix C introduces categorical probability, and in fact goes into *more detail than is necessary* for the course itself. In that sense, Appendix C is *entirely optional* and exists for those curious souls who wish to learn a little bit of categorical probability ahead of the main lectures.

Before going any further, we need to establish what, exactly, *computer intensive statistics* actually is. Clearly, it *is* statistics and this must be borne in mind throughout: the computer is a tool which we are using to resolve statistical problems as well as we are able to. When we can dispense with complicated computational procedures without sacrificing accuracy or power then we should probably do so.

Ultimately, statistics is all about using *mathematical models* to make sense of *data*. So what sort of statistics is *computer intensive*? There are several situations in which we are likely to find ourselves needing substantial computing power:

Large datasets. It is hard to overstate just how much data is being collected nowadays. Virtually every app on a smartphone or 'smart' device is constantly collecting usage data. In medicine and health, drug discovery is hugely driven by vast quantities of health data. Within the realm of AI, it is estimated that current language models are trained on 10-15 trillion tokens (equivalent to about 2000 times the entirety of English Wikipedia). And large scientific labs like CERN can be generating as much as 1,000 terabytes of data per day. Doing *anything* interesting with data on such scales is certainly computer intensive.

Large Models. We are also often in situations where the models we are fitting to our data are extremely large (e.g. in terms of numbers of parameters) or otherwise complex. Situations include:

[1] Apologies to those of you who are disappointed by this. There will be copious references to places which do.

- large, hierarchical Bayesian models with many levels of uncertainty and many latent variables;
- scientific models from which we can sample but which are so complicated that we can't write down the likelihood—think of models in climate science, numerically solved differential equation models in physics, etc.;
- models from deep learning and AI which are massively overparameterized and have a vast number of parameters.

Roughly speaking, everything which we will discuss can be thought of as being computer intensive for one of two reasons: because we want to deal with so much data (in some sense) that doing anything with it is difficult *or* because what we want to do is intrinsically complicated.

Acknowledgements These materials are closely based on those developed by previous module leaders, Dr Richard Everitt (University of Warwick), Dr Paul Jenkins (University of Warwick) and Prof. Adam Johansen (University of Warwick), in turn based on a course Prof. Johansen developed with Dr Ludger Evers (University of Glasgow).

2 Towards Computer Intensive Statistics

This chapter aims to introduce a few of the ideas which will be important in this module and to provide some pointers to directions which will be looked at during the week.

There are a great many books on simulation-based computational statistics; some good examples are those of Voss (2013) which provides a gentle introduction to simulation-based computational methods, Robert and Casella (2004) which takes a rigorous but approachable look at the area from a mathematical/statistical perspective and Liu (2001) which many with backgrounds in the natural sciences find to be very accessible. A more recent book which focusses on issues of scalability is Fearnhead et al. (2025).

2.1 Simulation and the Monte Carlo Method

Simulation will occupy most of our time in this module. The idea of drawing random variables from specified distributions and using the resulting ensemble of realisations to approximate quantities of interest is intoxicatingly powerful for a method which, at least in principle, is very simple.

Methods based around simulation are typically termed Monte Carlo[2] methods, following Metropolis and Ulam (1949). One characterization of such methods was given by Halton (1970):

> Representing the solution of a problem as a parameter of a hypothetical population, and using a random sequence of numbers to construct a sample of the population, from which statistical estimates of the parameter can be obtained.

In some ways this is *doing statistics backwards*: rather than taking a real sample (of data) and attempting to infer parameters of an underlying population using analytical techniques, we devise a representation of a quantity of interest as a parameter of a hypothetical population and then obtain, artificially, a sample from that population before using the properties of this sample as a proxy for those of the population itself.

Perhaps an example might make it clear how simple this approach really is. You may well have seen this example before, but try to see it as a prototype for a general approach to approximate computation.

[2]Metropolis (1987) has this explanation: "It was at that time that I suggested an obvious name for the statistical method — a suggestion not unrelated to the fact that Stan had an uncle who would borrow money from relatives because he 'just had to go to Monte Carlo.' The name seems to have endured."

Example 2.1 (Computing pi in the rain)**.** Suppose we want to obtain an approximation of π using a simple experiment. Assume that we are able to produce "uniform rain" on the square extending to ± 1 in two orthogonal directions, $[-1,1]^2 = [-1,1] \times [-1,1] = \{(x,y) : x \in [-1,1], y \in [-1,1]\}$, such that the probability of a raindrop falling into a region $\mathcal{R} \subseteq [-1,1]^2$ is proportional to the area of $\mathcal{R}$, but independent of the position of $\mathcal{R}$. It is easy to see that this is the case iff the two coordinates X, Y are independent realisations of uniform distributions on the interval $[-1,1]$ (in short $X, Y \overset{\text{iid}}{\sim} \mathsf{U}[-1,+1]$).

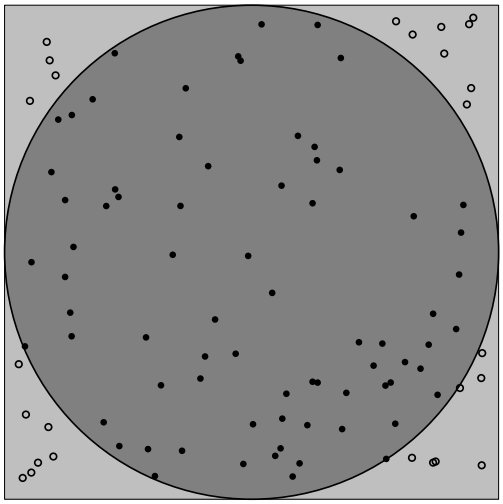

Figure 1: Illustration of the estimation π using uniform rain.

Now consider the probability that a raindrop falls into the unit circle (see Figure 1). It is

$$\mathbb{P}(\textit{drop within circle}) = \frac{\textit{area of the unit circle}}{\textit{area of the square}} = \frac{\iint\limits_{\{x^2+y^2\leq 1\}} 1 \, dxdy}{\iint\limits_{\{-1\leq x,y\leq 1\}} 1 \, dxdy} = \frac{\pi}{2 \cdot 2} = \frac{\pi}{4}.$$

In other words,

$$\pi = 4 \cdot \mathbb{P}(\textit{drop within circle}),$$

i.e. there is an expression for the desired quantity π as a function of a probability.

Of course we cannot compute $\mathbb{P}(\textit{drop within circle})$ without knowing π. However, we can estimate the probability using our raindrop experiment. If we observe n raindrops, then the number of raindrops M that fall inside the circle is a binomial random variable, $M \sim \mathsf{Bin}(n,p)$ with $p = \mathbb{P}(\textit{drop within circle})$.

Thus, if we observe that m raindrops fall within the circle, we approximate p using its maximum-likelihood estimate $\hat{p} = m/n$, and we can estimate π by $\hat{\pi} = 4\hat{p} = 4 \cdot \frac{m}{n}$. Assume we have observed, as in Figure 1, that 77 of the 100 raindrops were inside the circle. In this case, our estimate of π is $\hat{\pi} = 4 \times 77/100 = 3.08$, which is clearly some way from the truth.

However the ***strong law of large numbers*** *(Theorem B.2) guarantees that the estimator $\hat{\pi} = 4M/n$ converges almost surely to π. Figure 2 shows the estimate obtained after n iterations as a function of n for $n = 1, \ldots, 2000$. You can see that the estimate improves as n increases.*

We can assess the quality of our estimate by computing a confidence interval for π. As we have $X \sim \mathsf{Bin}(100,p)$, we can obtain a 95% confidence interval for p using a Normal approximation:

$$\left[0.77 - 1.96 \cdot \sqrt{\frac{0.77 \cdot (1-0.77)}{100}},\ 0.77 + 1.96 \cdot \sqrt{\frac{0.77 \cdot (1-0.77)}{100}}\right] = [0.6875,\ 0.8525],$$

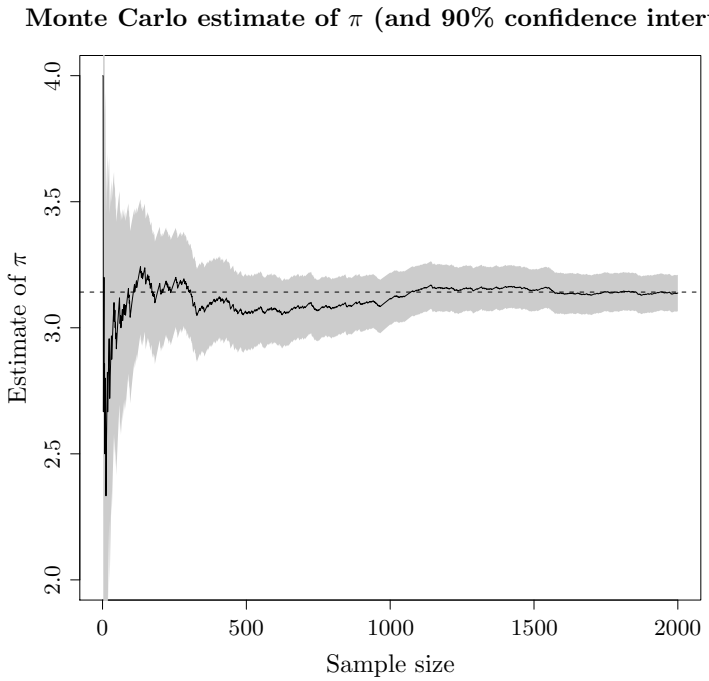

Figure 2: Estimate of π together with approximate confidence intervals resulting from the raindrop experiment. Notice the jagged evolution of the confidence interval estimate: this is something which must be borne in mind when using simulation, if we use uncertainty estimates based upon estimated values then the quality of the estimate will determine also the quality of the uncertainty estimate. We can dramatically underestimate our own uncertainty if we are not careful.

As our estimate of π is four times the estimate of p, we now also have a confidence interval for π which is simply $[2.750, 3.410]$.

In more general terms, let $\hat{\pi}_n = 4\hat{p}_n$ denote the estimate after having observed n raindrops. A $(1 - 2\alpha)$ confidence interval for π is

$$\left[\hat{\pi}_n - z_{1-\alpha}\sqrt{\frac{\hat{\pi}_n(4-\hat{\pi}_n)}{n}}, \hat{\pi}_n + z_{1-\alpha}\sqrt{\frac{\hat{\pi}_n(4-\hat{\pi}_n)}{n}}\right].$$

Recall the main steps of this process:

- We have written the quantity of interest (in our case π) as an expectation (a probability is a special case of an expectation as $\mathbb{P}(A) = \mathbb{E}[\mathbb{I}_A]$ where $\mathbb{I}_A(x)$ is the *indicator function* which takes value 1 if $x \in A$ and 0 otherwise).
- We have replaced this algebraic representation of the quantity of interest with a sample approximation. The strong law of large numbers guaranteed that the sample approximation converges to the algebraic representation, and thus to the quantity of interest. Furthermore, the central limit theorem (Theorem B.3) allows us to assess the speed of convergence.

Of course we'd never use a method like this one to estimate π; there are much faster ways of getting good estimates. Indeed, the rate of convergence here illustrates that these methods can be really computationally intensive. One major advantage of such methods, as we shall see, is that the *rate* of convergence of Monte Carlo estimates of expectations is independent of the dimension of the space on which the integral is defined, unlike more traditional approaches to numerical integration, and it is for hard problems in which other methods fail to produce any meaningful solution that simulation-based strategies are most useful.

2.2 Markov chains and Monte Carlo

Markov chains are objects with which you will become familiar during the APTS week, if you are not already. Roughly speaking, a Markov chain is a stochastic process for which the distribution of its future states is independent of its past given its current value. A more formal coverage of Markov chains will be provided by the *Applied Stochastic Processes* module, also taking place this APTS week! (In fact, during the week, we will actually utilise the recent perspective of *categorical probability*. This is discussed in the Appendix C.)

The *ergodic hypothesis* was originally the work of Boltzmann and was intended to provide a characterisation of the long term average behaviour of a thermodynamic system. It said, approximately, that the time averaged behaviour of the microscopic configuration of a system was the same as the instantaneous average over a hypothetical ensemble of systems prepared in a particular way. In modern terms we would think of that hypothetical ensemble as a way of describing a probability distribution and we could think of the ergodic hypothesis as telling us (assuming it to be true) that the long-time-average behaviour of the stochastic process describing the evolution of the system coincided with an expectation with respect to that probability distribution.

The previous paragraph appears to be hinting at something like a law of large numbers and, indeed, that's what a modern ergodic theorem describes. From a simulation point of view this is a tremendously powerful concept and there is an enormous literature on *Markov chain Monte Carlo* methods in which the trajectories of Markov chains are simulated for a long period of time and averages over these trajectories are calculated as a proxy for the expectation with respect to a particular distribution. A major contributing factor to the popularity and pre-eminence of these methods amongst computational statistics is that there exist recipes for the construction of Markov chains whose ergodic averages coincide with expectations with respect to essentially any distribution of interest, particularly the Metropolis (Metropolis et al. 1953) algorithm and its variants such as the Metropolis–Hastings algorithm (Hastings 1970) and the reversible jump algorithm

(Green 1995). For some well-behaved statistical problems, the so-called Gibbs sampler (Geman and Geman 1984) provides what may seem an even more automatic approach, but it does come at the cost of some additional human calculation.

During the course of the APTS week we'll see how to construct Markov chains to answer particular questions of interest and touch on all of the algorithms mentioned above.

There are vast numbers of resources which provide information on Markov chain Monte Carlo methods. I find that Robert and Casella (2004) covers this particular sub-field of Monte Carlo very well, but many other resources exist. For many years Gilks, Richardson, and Spiegelhalter (1996) has been something of a canonical reference, but has perhaps aged a little since its publication; perhaps less venerable but undoubtedly more up to date is Brooks et al. (2011) and even more recently, see Fearnhead et al. (2025), which focuses on issues of scalability.

2.3 Big Data is a Big Problem

Here's the description of a dataset from a piece of work in which previous module leader Paul Jenkins was involved (Chan, Jenkins, and Song 2012):

> We applied our method to samples from two populations of D. melanogaster (fruit flies): Raleigh, USA (RAL) and Gikongoro, Rwanda (RG). The RAL dataset consisted of the genomes of 37 inbred lines sequenced at a coverage of $\geq 10\times$ by the Drosophila Population Genomics Project. The RG dataset comprised 22 genomes from haploid embryos sequenced at a coverage of $\geq 25\times$ by the Drosophila Population Genomics Project 2... we were able to designate the ancestral allele in 1,755,040 of 2,475,674 high quality (quality score $Q \geq 30$) SNPs (single nucleotide polymorphisms) in the RAL sample (70.9%), and 2,213,312 out of 3,134,295 high quality SNPs in the RG sample (70.6%).

Each genome is summarised by its positions that differ from some other members of the sample ("SNPs"), so the data can be summarised by tables of size $37 \times 2,475,674$ (RAL) and $22 \times 3,134,295$ (RG). By the standards of modern statistics, these *aren't* particularly large data sets, yet managing, storing and performing inference for data sets on this scale required considerable thought and effort.

The need for tools which can deal efficiently with data sets of this magnitude (and, indeed, much larger ones) is one of the reasons that computer intensive statistics is important. It would not be feasible to compute the sample mean of a data set with a quarter of a million elements without making use of a computer; of course, performing meaningful inference typically requires much more sophisticated computations than that.

A question to think about: from your personal perspective: how big is a *large* data set and how big is an *enormous* data set?

2.4 Generative Modelling

In recent years, with the advent of generative AI, an entirely new paradigm for sampling has emerged. In traditional sampling problems – such as in Example 2.1 – the sampling distribution is known, and we simply need to generate samples, either through direct sampling, or using something like Markov chain Monte Carlo, which returns approximate samples.

In the context of generative modelling, the assumption is that we *begin with samples* $X_1, \ldots, X_n \sim \pi$, which are assumed to be i.i.d. from some *unknown* distribution π. For example, we might just have a large database of unlabelled cat images. The goal of generative modelling is then to *produce new samples* $X_1^*, \ldots, X_m^*$ which are also (approximately) distributed according to π. In other words, we want to generate new cat images. The crucial difficulty here is that we simply don't know what π is (it's the distribution of all possible cat images!) and so we have to learn everything from the samples themselves.

You may well have come across AI image generators, such as Stable Diffusion or DALL-E - these are all extremely powerful examples of generative modelling. Overall, generative modelling has exploded as a field in the last few years, and the field continues to move extremely fast. In terms of computer intensive statistics, these ideas have also found applications in simulation-based inference e.g. Sharrock et al. (2024), and we might touch on some of these towards the end of the course.

2.5 Warm-Up Exercises

Exercise 2.1. (Preliminary Simulation). *Familiarise yourself with the support provided by R for simple simulation tasks. In particular:*

1. *Generate a large sample of* $\mathsf{N}(3,7)$ *random variables (see* `rnorm`*); plot a histogram (*`hist` *with sensibly-chosen bins) of your sample and overlay a normal density upon it (see* `dnorm, lines`*).*
2. *Use* `sample` *to simulate the rolling of 1,000 dice; use your sample to estimate the average value obtained when rolling a standard die. Show how the estimate obtained using* n *dice behaves for* $n \in [0, 1000]$ *using a simple plot.*

Exercise 2.2. (Bootstrap Methods). *Imagine that you have a sample of 1,000 values from a population of unknown distribution (for our purposes, you can obtain such a sample using* `rnorm` *as in the previous question and pretending that the distribution is unknown).*

1. *Write code to repeat the following 1,000 times:*
   (a) *Sample 1,000 times with replacement from the original sample to obtain 1,000* ***resampled*** *sets of values.*
   (b) *Compute the sample mean of your resampled set.*
2. *You now have 1,000 sample means for resampled subsets of your data. Find the 5th and 95th percentile of this collection of resampled means.*
3. *How does this compare with a standard 90% confidence interval for the mean of a sample of size 1,000 from your chosen distribution?*
4. *Repeat the above using the median rather than mean.*
5. *Why* ***might*** *this be a useful technique? Note that we haven't done anything to justify the approach.*

Exercise 2.3. (Transformation Methods). *The* ***Box–Muller method*** *transforms a pair of uniformly-distributed random variables to obtain a pair of independent standard normal random variates. If*

$$U_1, U_2 \overset{iid}{\sim} \mathsf{U}[0,1],$$

and

$$\begin{aligned} X_1 &= \sqrt{-2\log(U_1)} \cdot \cos(2\pi U_2), \\ X_2 &= \sqrt{-2\log(U_1)} \cdot \sin(2\pi U_2), \end{aligned}$$

then $X_1, X_2 \overset{iid}{\sim} \mathsf{N}(0,1)$.

(a) *Write a function which takes as arguments two vectors* $(\mathbf{U}_1, \mathbf{U}_2)$ *of uniformly distributed random variables, and returns the two vectors* $(\mathbf{X}_1, \mathbf{X}_2)$ *obtained by applying the Box–Muller transform elementwise.*

(b) *The R function* `runif` *provides access to a 'pseudo-random number generator' (PRNG), which we'll discuss further during the module. Generate 10,000* $\mathsf{U}[0,1]$ *random variables using this function, and transform this vector to two vectors, each of 5,000 normal random variates.*

(c) *Check that the result from (b) is plausibly distributed as pairs of independent, standard normal random variables, by creating a scatter plot of your data.*

Exercise 2.4. (Simulating Markov chains). *Consider a simple board game in which players take it in turns to move around a circular board in which an annulus is divided into 40 segments. Players move by rolling two standard dice and moving their playing piece, clockwise around the annulus, the number of spaces indicated. For simplicity we'll neglect the game's other features.*

1. *All players begin in a common space. Write R code to simulate a player's position after three moves of the game and repeat this 1,000 or so times. Plot a histogram to show the distribution of player positions after three moves. Is this consistent with your expectations?*
2. *Now modify your code to simulate the sequence of spaces occupied by a player over their first 10,000 moves. Plot a histogram to show the occupancy of each of the forty spaces during these 10,000 moves. Are there any interesting features?*
3. *If a player's score increases by 1 every time they land on the first space (the starting space), 2 every time they land in the space after that and so on up to 40 for landing in the space immediately before that space then approximately what would be the long-run average number of points per move (use your simulation to obtain an approximate answer).*

A Inference and Estimators

Computer intensive statistics is still statistics, and it is important not to lose sight of that fact. It's necessary to focus on the core ideas of computer intensive statistics in this module due to the limited time available, but it is important to remember *why* we want to address the problems we consider.

One of the major tasks of computer intensive statistics is to provide (approximate) estimates in situations in which the estimators of interest are analytically intractable. This chapter provides a short reminder of some of the estimators of widespread use in statistics that will feature in this module.

Most if not all of this material was covered in much greater depth in the *Statistical Inference* module.

A.1 Inference and Some Common Estimators

Computationally intensive methods are extremely prevalent in Bayesian statistics and people often make the mistake of thinking the two are synonymous. This is not the case. Computer intensive methods can be very useful in other areas of statistics, including likelihood-based inference. With this in mind, it is useful to recall a number of common estimation tasks we might want to carry out. It is very likely that you've come across all of these things before; the particular character of our interest is that we shall seek to find approximate solutions to these estimation problems in settings in which they are not analytically tractable (either because the computation is apparently not possible even in abstract terms or because carrying it out would take many times the age of the universe).

A.1.1 Maximum Likelihood Estimates

Given a generative model for our data, $\boldsymbol{X} \sim f(\cdot \mid \boldsymbol{\theta})$, the likelihood is the function $L(\boldsymbol{\theta}; \boldsymbol{x})$ viewed as a function of the parameter vector, $\boldsymbol{\theta}$, with the *observed data*, $\boldsymbol{x}$, treated as fixed.

The *maximum likelihood estimator* is:

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} := \arg\max_{\theta} L(\boldsymbol{\theta}; \boldsymbol{x}).$$

It is common to work with the logarithm of the likelihood, $\ell(\boldsymbol{\theta}; \boldsymbol{x}) = \log L(\boldsymbol{\theta}; \boldsymbol{x})$ for numerical reasons. In particular, if $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $X_i \overset{\text{iid}}{\sim} f(\cdot|\boldsymbol{\theta})$ then:

$$L(\boldsymbol{\theta}; \boldsymbol{x}) = \prod_{i=1}^{n} f(x_i \mid \boldsymbol{\theta}), \quad \ell(\boldsymbol{\theta}; \boldsymbol{x}) = \sum_{i=1}^{n} \log f(x_i \mid \boldsymbol{\theta}),$$

and it is typically much easier to work with ℓ than L. By the strict monotonicity of the logarithm (and non-negativity of the likelihood) it is clear that:

$$\arg\max_{\theta} L(\boldsymbol{\theta}; \boldsymbol{x}) = \arg\max_{\theta} \ell(\boldsymbol{\theta}; \boldsymbol{x}).$$

For complex models, obtaining this maximiser analytically can be impossible; we'll see that there are (at least approximate) computational solutions to this problem during this module.

A.1.2 Confidence Intervals

A confidence interval is a *random* set which will contain the true value of the parameter with a specified probability with respect to replication of the sampling experiment.

If we are interested in a (real-valued) parameter θ and a density $f(\boldsymbol{x}; \theta)$ describing a data-generating process then we seek random variables $L(\boldsymbol{X})$ and $U(\boldsymbol{X})$ such that $\mathbb{P}(L(\boldsymbol{X}) \leq \theta \leq U(\boldsymbol{X})) = 1 - \alpha$ for some *confidence level* α. Note that θ is not treated as random here, but as fixed and unknown; it is $L(\boldsymbol{X})$ and $U(\boldsymbol{X})$ that are random and the probability is with respect to their distribution under repeated realisations of the experiment which realises the random variable $\boldsymbol{X}$ describing the data.

A *level α confidence interval for θ*, then, is a random interval $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ which would contain the true parameter a proportion (approximately / exactly) α of the time if we carried out the whole procedure (a very large number of / infinitely many) times.

A.1.3 Hypothesis Tests

Closely related to the notion of a confidence interval is the *hypothesis test*. In order to distinguish between two possible explanations for observed data, a default scenario H_0 termed the *null hypothesis* and an alternative H_1, this procedure seeks to *reject* H_0 when the data is in an appropriate sense unlikely to have arisen if H_0 is true and *not to reject* H_0 otherwise.

More precisely, given a test statistic $T(\boldsymbol{X})$, we seek a set of values C_α such that

$$\mathbb{P}(T(\boldsymbol{X}) \in C_\alpha \mid H_0 \text{ is true}) = \alpha, \quad \mathbb{P}(T(\boldsymbol{X}) \in C_\alpha \mid H_1 \text{ is true}) > \alpha.$$

Often C_α is obtained as the complement of an interval of values which are likely under the null hypothesis (viewed relatively, contrasting with the plausibility of those values under the particular alternative hypothesis). See Figure 3 for an illustration.

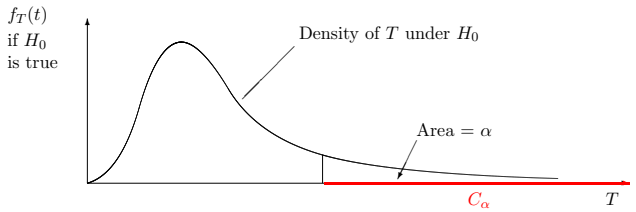

Figure 3: The critical or rejection region associated with a particular test; $C_\alpha = [c, \infty)$ in this case.

Computing the distribution of $T(\boldsymbol{X})$ under H_0 can be difficult (essentially impossible) in complicated situations. We will see that there are computational solutions to this problem.

A.1.4 Bayesian Point Estimates

In Bayesian statistics we summarise the state of knowledge about unknown variables using a probability distribution. The *prior* distribution specifies the state of knowledge about an unknown parameter vector θ prior to the current experiment; again a generative model $f(\boldsymbol{x} \mid \theta)$ then describes the distribution of the data given a particular value for this parameter. Combining these according to the standard probability calculus gives rise to the celebrated *Bayes rule*:

$$p(\theta \mid \boldsymbol{x}) = \frac{p(\theta) f(\boldsymbol{x} \mid \theta)}{p(\boldsymbol{x})},$$

which connects the *posterior* distribution $p(\theta \mid \boldsymbol{x})$ (summarising knowledge about the parameter vector after the assimilation of the current data) to those quantities. Here, $p(\boldsymbol{x}) = \int p(\theta) f(\boldsymbol{x} \mid \theta) d\theta$ is sometimes known as the *evidence* or *marginal likelihood* and allows the comparison of different models: it tells us how likely particular data is given a particular model (and given a prior distribution over the parameters of that model).

There are some estimates which are very commonly used (posterior mean and median, for example) but in principle one should proceed by specifying a loss (or cost) function $C : \Theta \times \Theta \to [0, \infty)$ where $C(\theta, \vartheta)$ specifies the cost of estimating a value of θ when the true parameter value is ϑ. See Robert (2001) for an in depth discussion of this approach. We'll settle in this module for obtaining approximations of these estimates by computational methods.

A.1.5 Credible Intervals

Bayesian interval estimates are especially simple. A *credible interval* of level α for a real-valued parameter θ is an interval $[L(\boldsymbol{x}), U(\boldsymbol{x})]$ that contains the parameter θ with probability α conditional upon having observed the particular data $\boldsymbol{x}$, i.e. $\mathbb{P}(\theta \in [L(\boldsymbol{x}), U(\boldsymbol{x})] \mid \boldsymbol{X} = \boldsymbol{x}) = \alpha$. As in the Bayesian paradigm, θ is itself a random variable, so this probabilistic statement makes sense and credible intervals admit a much simpler interpretation than the confidence intervals which they superficially resemble.

We'll see that credible intervals can be obtained from essentially the same computational methods as Bayesian point estimates.

A.2 Variability of Estimators and Uncertainty Quantification

A theme that turns out to be important in a number of forms in computer intensive statistics is the variability of estimate and the quantification of uncertainty. In classical statistics, the sampling distribution of an estimator is often used to provide some measure of uncertainty, perhaps via confidence intervals. The sampling distribution of a statistic $T(\boldsymbol{X})$ is simply the distribution which it has under repeated sampling of the data itself $\boldsymbol{X}$ from the model. In Bayesian statistics, the posterior distribution itself, $p(\theta \mid \boldsymbol{x})$, summarises the uncertainty we have about the value of any parameter after incorporating the information contained in the data we have available. Both of these are things which we may wish to estimate *using* computer intensive statistics.

It is important to distinguish between the sampling distribution of a statistic or the posterior variance, which summarise in different ways the degree of uncertainty which must accompany any point estimate, and additional variability introduced by the procedure used to approximate an estimator. In particular, we will see that we often introduce additional auxiliary stochasticity during the computational procedures we use; this is undesirable and we seek to minimise it and to mitigate any influence it may have upon our estimation.

A.3 Warm-Up Exercises

The main purpose of the following is to show some places in which it quickly becomes difficult to proceed analytically, but for which computational methods might work well, and to highlight the types of quantities which we will want to be able to compute or approximate in this module.

Exercise A.1. *What are the maximum likelihood estimators for the parameters in the following situations:*

(a) $X_1, \ldots, X_n \overset{iid}{\sim} \mathsf{N}\left(\mu, \sigma^2\right)$ *where* μ *and* σ^2 *are unknown;*

(b) $X_1, \ldots, X_n \overset{iid}{\sim} \mathsf{U}[0, \theta]$ *where* θ *is unknown.*

(c) $X_1, \ldots, X_n \overset{iid}{\sim} f(x; \mu_1, \mu_2, w)$ *where* $f(x; \mu_1, \mu_2, w) := w\mathsf{N}(x; \mu_1, 1) + (1-w)\mathsf{N}(x; \mu_2, 1)$*, with* $w \in (0, 1)$ *and* $\mu_1, \mu_2 \in \mathbb{R}$*?*

Exercise A.2. *Find the estimators which minimise the posterior expected loss for the following loss functions for* ***continuous*** *parameters:*

(a) *Squared-error (or quadratic) loss:* $C(\theta, \vartheta) = (\theta - \vartheta)^2$.

(b) *Absolute-error loss:* $C(\theta, \vartheta) = |\theta - \vartheta|$.

and for ***discrete*** *parameters:*

1. *Squared-error (or quadratic) loss:* $C(\theta, \vartheta) = (\theta - \vartheta)^2$.
2. *Zero-one loss:*

$$C(\theta, \vartheta) = \begin{cases} 0 & \text{if } \theta = \vartheta \\ 1 & \text{otherwise.} \end{cases}$$

B Convergence

Recall that a sequence of real numbers, $x_1, x_2, \ldots$ is said to *converge to a limit* x, written $\lim_{n\to\infty} x_n = x$ or $x_n \to x$, if for every $\epsilon > 0$ there exists some n_0 such that, for all $n > n_0$ we have that $|x_n - x| < \epsilon$.

When we deal instead with stochastic objects—random variables or empirical distributions, for example—we need to expand upon this idea somewhat and there are a number of natural extensions. Below we recall some common stochastic notions of convergence together with some key theorems about these types of convergence. These are of great importance in statistics in general and computational statistics in particular.

Here we consider only real-valued random variables. In computational statistics it is often necessary to consider the convergence of more complicated stochastic quantities, but in the current module a qualitative understanding of the following notions of convergence should be more than sufficient and so we avoid technical details. There are many excellent books on these topics; Shiryaev (1995) provides a rigorous but accessible treatment.

B.1 A Word on Probability

We won't see technical (measure theoretic) probability in this module—we can manage without it for our purposes. In fact, we will even consider an alternative approach (categorical probability - Appendix C which circumvents it entirely). There are a few places where this may cause us a slightly loss of generality, but these will be highlighted.

When we talk about probability we assume that there is some underlying *sample space*, Ω, from which exactly one outcome occurs every time our *experiment* is realised: an *elementary event*, $\omega \in \Omega$. A *random variable* can be thought of as a measurement which we could make when the experiment is carried out and can be modelled mathematically as a function which maps the sample space to the real numbers, $X : \Omega \to \mathbb{R}$ (see Figure 4).

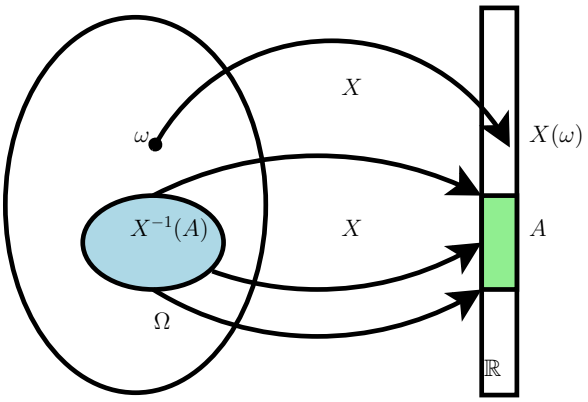

Figure 4: Random variables as functions from the sample space, Ω, to $\mathbb{R}$. If ω occurs in an experiment then X will be measured as having the value $X(\omega)$. If we want to know whether X takes a value in the set A (the green rectangle) then we need to check whether ω takes a value which X maps into A, this set is often denoted $X^{-1}(A)$ and corresponds to the blue ellipse here.

When we talk about the probability of some event $B \subseteq \Omega$ we mean exactly $\mathbb{P}(B) := \mathbb{P}(\omega \in B)$, i.e. the probability that the elementary outcome which occurs is contained within B. When we talk about the probability of a random variable, X, taking a value in a set, e.g. $\mathbb{P}(X \in A)$ we really mean the probability that the elementary outcome that occurs is such that X takes a value in A: $\mathbb{P}(X \in A) := \mathbb{P}(\{\omega : X(\omega) \in A\}) = \mathbb{P}(\omega \in X^{-1}(A))$ where $X^{-1}(A)$ denotes the *pre-image* of A under X, i.e. the collection of all points in Ω which are mapped into A by X: $X^{-1}(A) = \{\omega : X(w) \in A\}$.

B.2 Convergence In Probability

A sequence of random variables $X_1, X_2, \ldots$ is said to *converge in probability* to a limiting value x, written $\lim_{n\to\infty} X_n = x$ (in probability) or $X_n \xrightarrow{p} x$, if for every $\epsilon > 0$ the probability that X_n is further from x than ϵ converges to zero, i.e.

$$\lim_{n\to\infty} \mathbb{P}\left(|X_n - x| > \epsilon\right) = 0.$$

The first theoretical result of interest tells us that in this sense, averages obtained from simple samples will converge to the population average.

Theorem B.1. *(Weak Law of Large Numbers; see, for example, Shiryaev (1995), p325).* Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables and let $\varphi : \mathbb{R} \to \mathbb{R}$ be a function of interest. If $\mathbb{E}\left[|\varphi(X_1)|\right] < \infty$ then, as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} \varphi(X_i) \xrightarrow{p} \mathbb{E}[\varphi(X_1)].$$

B.3 Almost Sure Convergence

A sequence of random variables $X_1, X_2, \ldots$ is said to *converge almost surely (a.s.)* or *converge with probability 1* to a limiting value x, written $\lim_{n\to\infty} X_n = x$ (a.s.) or $X_n \xrightarrow{a.s.} x$, if $\mathbb{P}\left(X_n \to x\right) = \mathbb{P}\left(\left\{\omega : X_n(\omega) \to x\right\}\right) = 1$.

Almost sure convergence is strictly stronger than convergence in probability. We can, however, be sure that under weak assumptions the average from a simple random sample will convergence almost surely to the underlying population average.

Theorem B.2. *(Strong Law of Large Numbers; see, for example, Shiryaev (1995), p391).* Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables and let $\varphi : \mathbb{R} \to \mathbb{R}$ with $\mathbb{E}\left[|\varphi(X_1)|\right] < \infty$. The sample average of φ converges, as $n \to \infty$, to its expectation under the common distribution of the X_i with probability 1:

$$\frac{1}{n}\sum_{i=1}^{n} \varphi(X_i) \overset{a.s.}{\to} \mathbb{E}\left[\varphi(X_1)\right].$$

B.4 Some Ideas Related to Convergence of Distributions

B.4.1 Convergence In Distribution

A sequence of random variables $X_1, X_2, \ldots$ is said to converge in distribution to another random variable X if, for every continuous bounded function $\varphi : \mathbb{R} \to \mathbb{R}$ we have that $\mathbb{E}\left[\varphi(X_i)\right] \to \mathbb{E}\left[\varphi(X)\right]$. If we allow F_i to denote the distribution function of X_i and F that of X then this mode of convergence is equivalent to the *pointwise* convergence of F_i to F (except at an at most countable collection of points of discontinuity).

Although this may seem an esoteric idea, it's really just telling us that the distribution of the sequence of random variables becomes arbitrary close to that of X, eventually. Any statistician is familiar with the following example of convergence in distribution.

Theorem B.3. *(Central Limit Theorem; see Shao (1999), Corollary 1.2, for example).* Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ be independent and identically distributed k-dimensional random vectors (i.e. random elements in $\mathbb{R}^k$ which can be viewed as a vector of K random variables) with finite covariance matrix Σ, then as $n \to \infty$,

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i - \mathbb{E}\left[\boldsymbol{X}_1\right]\right] \overset{\mathcal{D}}{\to} \mathsf{N}\left(\mathbf{0}, \boldsymbol{\Sigma}\right),$$

where $\mathbf{0}$ denotes the zero element of $\mathbb{R}^k$.

B.4.2 Glivenko–Cantelli Theorems

The following result isn't cast as convergence in distribution but it does tell us something very closely related and so it is included here. This result is perhaps slightly less widely known than those mentioned above, but it is tremendously informative for many of the methods which we will consider in this module.

Theorem B.4. *(Glivenko–Cantelli; see Athreya (2003) for a self-contained proof).* Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables with distribution function F. Let $F_n(x)$ denote the **empirical distribution functions** associated with the first n elements in this sequence, i.e. let

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{(-\infty, X_i]}(x),$$

where

$$\mathbb{I}_{(-\infty, X_i]}(x) = \begin{cases} 1 & \text{if } x \in (-\infty, X_i] \\ 0 & \text{otherwise.} \end{cases}$$

Then, as $n \to \infty$,

$$\sup_x |F_n(x) - F(x)| \overset{a.s.}{\to} 0.$$

This tells us that if we construct a probability distribution by placing a mass of $1/n$ at the location of every one of a sample of n independent and identically distributed replicates of a random variable with a given distribution function then, for large enough samples, that *empirical* distribution function converges uniformly to the underlying distribution function. Which tells us in a precise sense something which we might intuitively have believed: if we replace the original probability distribution with that obtained from a large enough sample then for most practical purposes we will obtain a good approximation.

B.5 Warm-Up Exercises

If you feel like you could do with reminding yourself how these ideas of stochastic convergence work then you might like to have a go at the following; if you already *know* how to answer them then don't waste your time and, similarly, if you have no interest in such things then it won't be essential to the lectures for this module that you have worked these out.

Exercise B.1. *Given an example of a sequence of random variables which:*

(a) *converge to a limit with probability one;*

(b) *converge to a limit in probability but* ***not*** *almost surely; and*

(c) *converge to a limit in distribution but* ***not*** *in probability.*

C Categorical Probability

Measure-theoretic probability (as touched on in Appendix B) is the foundation for probability theory, as is currently taught at the undergraduate level. However, the fact that we won't really make use of it in this course should signal something: for those of us interested in statistics, *measure-theoretic probability is not really crucial*.

This appendix introduces a recently proposed alternative perspective on probability Fritz (2020), Cho and Jacobs (2019), based on *category theory*. We will make use of categorical probability during the lectures to streamline several proofs. In order to follow those proofs, it is not necessary to have any prior understanding of categorical probability or category theory, so in that sense this appendix is *entirely optional*. If, having read this section, you feel you who want to learn a little more category theory, I highly recommend Perrone (2024).

Traditionally, probability theory is built on measure theory as outlined in Appendix B. This is a rigorous and proper foundation for probability, but not without its deficiencies. In particular, for statistics and machine learning, it's extremely rare that one requires the 'low-level' machinery of measure theory. We virtually never construct explicitly the underlying sample space Ω or σ-algebra $\mathcal{F}$, and even less define our random variables explicitly as measurable mappings $X : \Omega \to \mathbb{R}$. In practice (say, when coding), we simply generate random variables whenever we need them and transform them and work with them without every worrying about issues such as measurability. (After all, our computers are finite machines and so cannot do anything non-measurable, which would require an uncountable number of actions!)

The alternative perspective of categorical probability instead starts with *composition*. The primary elements become (generalised notions of) *Markov kernels*, which we can think of as random functions $P : \mathcal{X} \to \mathcal{Y}$. In other words, the random function P takes as input some $x \in \mathcal{X}$, and will return some $Y \in \mathcal{Y}$, but crucially this output Y is random. The key properties, then, are how these random functions *compose with one another*. In order to make this more precise, we have to start with the fundamental definition.

Definition C.1 (Category)**.** A category C consists of a collection of *objects*, denoted $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \ldots$ and a collection of *morphisms* (or *arrows*) between objects, denoted $f : \mathcal{X} \to \mathcal{Y}$, such that:

- given $g : \mathcal{Y} \to \mathcal{Z}$, we can form the *composite* morphism $g \circ f : \mathcal{X} \to \mathcal{Z}$;
- this composition is *associative*: given $h : \mathcal{Z} \to \mathcal{W}$, we have $(h \circ g) \circ f = h \circ (g \circ f)$;
- for every $\mathcal{X}$, there is an *identity* morphism $\mathrm{id}_{\mathcal{X}} : \mathcal{X} \to \mathcal{X}$, satisfying $f \circ \mathrm{id}_{\mathcal{X}} = f$ and $\mathrm{id}_{\mathcal{X}} \circ f' = f'$ for any $f' : \mathcal{Y} \to \mathcal{X}$.

Let's see some examples of Definition C.1, which are relevant for categorical probability.

Example C.1 (Examples of categories)**.**

1. **Meas**: objects are measurable spaces $(\mathcal{X}, \Sigma_{\mathcal{X}})$, morphisms are measurable functions $f : \mathcal{X} \to \mathcal{Y}$. Composition is function composition and the identities are the functions $x \mapsto x$.
2. **Stoch**: objects also measurable spaces, but morphisms are *Markov kernels*[3] $P : \mathcal{X} \to \mathcal{Y}$. Given kernel $Q : \mathcal{Y} \to \mathcal{Z}$, these are composed via Chapman–Kolmogorov:

$$Q \circ P(x, B) = \int_{\mathcal{Y}} P(x, \mathrm{d}y) Q(y, B).$$

   The identities are the kernels $x \mapsto \delta_x$.
3. **BorelStoch**: this is the same as **Stoch**, except as objects we only consider *standard Borel*[4] measurable spaces. Virtually any space used in practice in statistics is a standard Borel space (e.g. finite spaces, $\mathbb{R}^d$, etc). This additional regularity will be very useful when we later discuss *conditional distributions*.

C.1 The Giry functor

This subsection will introduce the Giry functor, which connects the categories **Meas** and **Stoch** previously introduced. This section will assume you are comfortable with measure-theoretic probability, so again is entirely optional. (We will not be making use of this section at all during the APTS week.)

Definition C.2 (Functor)**.** Let C, D be categories. A functor $F : \mathsf{C} \to \mathsf{D}$ is a mapping between the categories which preserves the categorical structure:

- For each object $\mathcal{X}$ in C, $F\mathcal{X}$ is an object in D.
- For each morphism $f : \mathcal{X} \to \mathcal{Y}$ in C, $Ff : F\mathcal{X} \to F\mathcal{Y}$ is a morphism in D. (Note we use the same notation to denote action on objects and morphisms - it is clear from the context which we mean.)
- Given another $g : \mathcal{Y} \to \mathcal{Z}$ in C, we have

$$F(g \circ_{\mathsf{C}} f) = Fg \circ_{\mathsf{D}} Ff,$$

  where we have made explicit which category the composition is occuring in.
- For any object $\mathcal{X}$ in C, $F\mathrm{id}_{\mathcal{X}} = \mathrm{id}_{F\mathcal{X}}$.

We now define a functor $\mathcal{P} : \mathbf{Meas} \to \mathbf{Meas}$ known as the *Giry functor*, Giry (1981). Following the definition of a functor, we need to specify its action on both objects and morphisms and check the remaining functor axioms.

So let's start with its action on objects. Let $\mathcal{X}$ be an object in **Meas**; in other words, $\mathcal{X}$ is a set equipped with a σ-algebra $\Sigma_{\mathcal{X}}$. We define $\mathcal{P}\mathcal{X}$ to be

$$\mathcal{P}\mathcal{X} = \{\mu : \mu \text{ is a probability measure on } \mathcal{X}\},$$

and we equip this set with the following σ-algebra:

$$\sigma(\mu \mapsto \mu(A) : A \in \Sigma_{\mathcal{X}}).$$

We need to do this since the objects of **Meas** are measurable spaces, not just sets. The intuition behind this σ-algebra is that it's the minimal one so that evaluating probabilities is measurable.

Now let's define the action on morphisms; so let $f : \mathcal{X} \to \mathcal{Y}$ be a morphism in **Meas**. In other words, it's a measurable function. We need to define a new morphism in **Meas**,

$$\mathcal{P}f : \mathcal{P}\mathcal{X} \to \mathcal{P}\mathcal{Y}.$$

[3] That is, $A \mapsto P(x, A)$ is a probability measure and $x \mapsto P(x, A)$ is measurable.

[4] Also known as Polish spaces: a topological space metrized by a complete and separate metric, equipped with its Borel σ-algebra.

In other words, this new function $\mathcal{P}f$ takes as input a probability measure μ on $\mathcal{X}$, and has to output some probability measure ν on $\mathcal{Y}$. There is a canonical choice for this, the *pushforward measure*: given such a μ, define ν to be $\mu^{\#}f$, the pushforward of μ under f:

$$\mu^{\#}f(B) = \mu(f^{-1}(B)), \quad B \in \Sigma_{\mathcal{Y}}.$$

In fact, this choice does indeed define a valid functor. The remaining details to check are routine (albeit a little tedious), and are left as an exercise:

Exercise C.1.

1. Check that the function $\mathcal{P}f : \mathcal{P}\mathcal{X} \to \mathcal{P}\mathcal{Y}$ defined above is in fact a *measurable* function. Hint: consider sets $M \subset \mathcal{P}\mathcal{Y}$ of the form $M = \{\nu : a < \nu(B) < b\}$ for $0 \leq a < b \leq 1$ and $B \in \Sigma_{\mathcal{Y}}$ and check $\mathcal{P}f^{-1}(M)$ is measurable.
2. Show that $\mathcal{P}$ preserves composition: $\mathcal{P}(g \circ f) = \mathcal{P}g \circ \mathcal{P}f$.
3. Show that $\mathcal{P}$ preserves identities: $\mathcal{P}\mathrm{id}_{\mathcal{X}} = \mathrm{id}_{\mathcal{P}\mathcal{X}}$.

In fact, the Giry functor is an example of a *monad*, which is another fundamental notion from category theory. And then, you can show that the category **Stoch** is precisely the *Kleisli category* arising from the Giry monad. I will not go into these details here (although the construction is extremely beautiful!) - I would refer you to the notes Perrone (2024), or feel free to get in touch with me!

References

Athreya, K. 2003. "A Simple Proof of the Glivenko–Cantelli Theorem." Technical Report. Cornell University Operations Research; Industrial Engineering.

Brooks, S., A. Gelman, G. L. Jones, and X.-L. Meng, eds. 2011. *Handbook of Markov Chain Monte Carlo*. CRC Press.

Chan, A. H., P. A. Jenkins, and Y. S. Song. 2012. "Genome-Wide Fine-Scale Recombination Rate Variation in *Drosophila Melanogaster*." *PLoS Genetics* 8 (12): e1003090.

Cho, Kenta, and Bart Jacobs. 2019. "Disintegration and Bayesian inversion via string diagrams." *Mathematical Structures in Computer Science* 29 (7): 938–71. https://doi.org/10.1017/S0960129518000488.

Fearnhead, Paul, Christopher Nemeth, Chris J. Oates, and Chris Sherlock. 2025. *Scalable Monte Carlo for Bayesian Learning*. Cambridge University Press. https://doi.org/10.1017/9781009288460.

Fritz, Tobias. 2020. "A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics." *Advances in Mathematics* 370: 107239. https://doi.org/10.1016/j.aim.2020.107239.

Geman, S., and D. Geman. 1984. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6 (6): 721–41.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter, eds. 1996. *Markov Chain Monte Carlo in Practice*. First. Chapman; Hall.

Giry, Michèle. 1981. "A categorical approach to probability theory." In *Categorical Aspects of Topology and Analysis*, edited by B. Banaschewski, 915:68–85. Lecture Notes in Mathematics. https://link.springer.com/book/10.1007/BFb0092866.

Green, P. J. 1995. "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination." *Biometrika* 82: 711–32.

Halton, J. H. 1970. "A Retrospective and Prospective Survey of the Monte Carlo Method." *SIAM Review* 12 (1): 1–63.

Hastings, W. K. 1970. "Monte Carlo Sampling Methods Using Markov Chains and Their Applications." *Biometrika* 52: 97–109.

Liu, J. S. 2001. *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics. New York: Springer Verlag.

Metropolis, N. 1987. "The Beginnings of the Monte Carlo Method." *Los Alamos Science* 15: 125–30. http://library.lanl.gov/cgi-bin/getfile?number15.htm.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, and A. H. Teller. 1953. "Equation of State Calculations by Fast Computing Machines." *Journal of Chemical Physics* 21: 1087–92.

Metropolis, N., and S. Ulam. 1949. "The Monte Carlo Method." *Journal of the American Statistical Association* 44 (247): 335–41. http://links.jstor.org/sici?sici=0162-1459%28194909%2944%3A247%3C335%3ATMCM%3E2.0.CO%3B2-3.

Perrone, Paolo. 2024. *Starting Category Theory*. Available at https://arxiv.org/abs/1912.10642; World Scientific Publishing Co. https://doi.org/10.1142/13670.

R Core Team. 2013. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Robert, C. P. 2001. *The Bayesian Choice*. 2nd ed. Springer Texts in Statistics. New York: Springer Verlag.

Robert, C. P., and G. Casella. 2004. *Monte Carlo Statistical Methods*. Second. New York: Springer Verlag.

Shao, J. 1999. *Mathematical Statistics*. Springer Texts in Statistics. Springer.

Sharrock, Louis, Jack Simons, Song Liu, and Mark Beaumont. 2024. "Sequential Neural Score Estimation: Likelihood-Free Inference with Conditional Score Based Diffusion Models." In *Proceedings of the 41st International Conference on Machine Learning*, 235:44565–602. PMLR. https://doi.org/10.48550/arxiv.2210.04872.

Shiryaev, A. N. 1995. *Probability*. Second. Graduate Texts in Mathematics 95. New York: Springer Verlag.

Voss, J. 2013. *An Introduction to Statistical Computing: A Simulation-Based Approach*. Wiley.