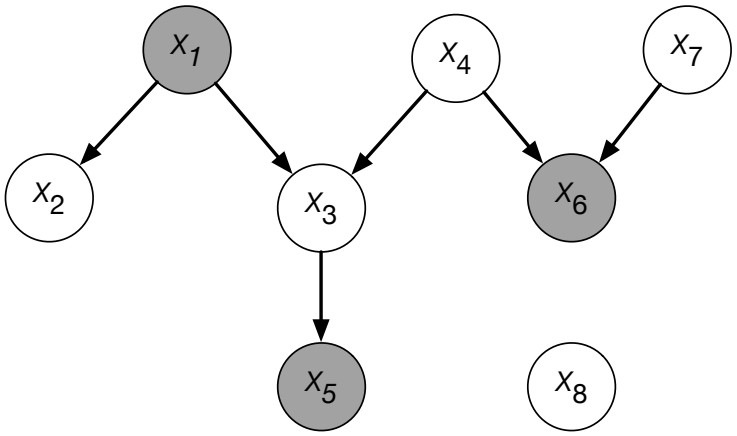
Probability, illustrated

Alexandre Bouchard-Côté



Contents

1 Motivations **6**

2 Foundations **7**

2.1 The basic vocabulary of probability 7
2.2 Motivating the axioms of probability 10
2.3 The axioms of probability 13
2.4 Some basic properties 14
2.5 Probability spaces as models 14
2.5.1 Coin tosses 14
2.5.2 Reliability 14
2.6 Statistical models 15
2.7 Simple examples of σ-algebra 15
2.8 More interesting examples of σ-algebra via generation 16
2.9 Simple examples of probability measures 17
2.10 Computations for discrete models 18
2.11 Exercise set 1 19
2.12 Solutions for exercise set 1 19
2.12.1 Basic properties 19
2.12.2 Reliability problem 20
2.12.3 Discrete computation 20
2.13 Random variables 22
2.14 Compositions of random variables 24
2.15 The graph of a random variable 25
2.16 The probabilist's event notation 25
2.17 Constant random variables 26
2.18 Indicator random variables 26
2.19 Modelling randomized algorithms as random variables 26
2.20 Distribution of a random variable 28
2.21 Cumulative distribution function 28
2.22 Equality in distribution 29
2.23 Densities (first definition) 30
2.24 Limit properties of probability measures 30
2.25 Limit properties of CDFs 31
2.26 Building random variables with a prescribed CDF 32
2.27 σ-algebra, revisited 33
2.28 Exercise set 2 33
2.29 Solutions for exercise set 2 34

3 Integration and expectation **36**
3.1 Overview . . . 36
3.2 Notation, inputs and outputs . . . 36
3.3 Generalizing the notion of the "area" of a "rectangle" . . . 37
3.4 Area under the graph of "simple functions" . . . 38
3.5 Area under graph of non-negative random variables . . . 38
3.6 Algorithmic construction . . . 39
3.7 Proving tool: simple function approximation + MCT . . . 40
3.8 Integrals of random variables taking negative values . . . 40
3.9 Integrals with respect to a measure . . . 41
3.10 More on exchanging limits and integrals . . . 41
3.11 Measure zero sets and almost sure statements . . . 42
3.12 Convexity and integration . . . 42
3.13 Markov's inequality and its friends . . . 44
3.14 Rewriting events involving equalities . . . 45
3.15 Rewriting events involving inequalities . . . 46
3.16 Bounding events . . . 46

4 Independence **46**
4.1 More than one random variables (random vectors) . . . 46
4.2 Distribution, CDF and density of a random vector . . . 47
4.3 Determining classes . . . 48
4.4 Independence of random variables . . . 49
4.5 Chernoff's bound, continued . . . 50
4.6 Exercise set 3 . . . 50
4.7 Solutions for exercise set 3 . . . 50

5 Computing expectations in practice **52**
5.1 Computing integrals using calculus . . . 52
5.2 Computing expectations using the distribution of the random variable . . . 53
5.3 How probability spaces and random variables are constructed in practice . . . 55
5.4 Computing expectations using densities . . . 55
5.5 Computing the expectation of a function of independent random variables . . . 56
5.6 Declaring independent random variables . . . 58
5.7 Computing expectation using the cumulative distribution function 58
5.8 Transformations of random variables and random vectors . . . 59

6 Asymptotics **60**
6.1 Infinitely often and eventually . . . 60
6.2 Borel-Cantelli (BC) lemma 1 . . . 61
6.3 Weak law of large number (WLLN) . . . 64
6.4 Convergence in probability . . . 65
6.5 Convergence almost surely . . . 66

6.6 Toward Central Limit Theorems . . . 67
6.7 Exact distribution of sums of random variables . . . 67
6.8 Interlude: exchanging the order of integration and differentiation 69
6.9 CLT: numerical exploration and intuition . . . 70
6.10 Basic CLT . . . 73
6.11 Exercise set 4 . . . 73
6.12 Solutions for exercise set 4 . . . 73
6.13 Types of convergence: big picture . . . 75
6.14 Weak convergence . . . 76
6.15 Overview of some properties of convergence of r.v.'s . . . 77
6.16 Towards a proof of the CLT: generating functions . . . 78
6.17 Probability generating functions . . . 79
6.18 Moment generating function . . . 81
6.19 Characteristic function . . . 81
6.20 Further properties of characteristic functions . . . 82
6.21 Proof of basic CLT . . . 84
6.22 CLT: multivariate version . . . 85
6.23 CLT: self-centered version . . . 86
6.24 Delta method . . . 87
6.25 LLNs and CLTs under relaxed assumptions . . . 88
6.26 Convergence in L^p . . . 88

7 Poisson theory **89**
7.1 Poisson convergence . . . 89
7.2 Poisson processes: motivation, definition and construction . . . 90
7.3 Poisson process with constant intensity on the real line . . . 94
7.4 Superposition . . . 95
7.5 Thinning . . . 95
7.6 Mapping . . . 96
7.7 Compound PP . . . 96

8 Conditioning **97**
8.1 Background: σ-algebra and information . . . 97
8.2 Conditioning on an event . . . 98
8.3 Conditioning on a random variable . . . 98
8.4 The law of total probability and expectation . . . 99
8.5 Key properties . . . 102
8.6 Equivalence of the two definitions . . . 103
8.7 The Bayes estimator (a special case) . . . 103
8.8 Geometric view of expectation and further properties . . . 104
8.9 Geometric view: more details on triangle inequality . . . 105
8.10 Conditional independence . . . 106
8.11 Directed graphical models . . . 106
8.12 Establishing (conditional) independence relations using directed graphical models . . . 107

9 Markov chains **109**
9.1 Basic definitions and examples 109
9.2 Representation under the homogeneity condition 110
9.3 First connection with linear algebra: Chapman-Kolmogorov equation 111
9.4 Hitting probabilities 111
9.5 Asymptotic behavior: overview 112
9.6 Law of large number for Markov chains 113
9.7 Extension to countably infinite spaces 115
9.8 Convergence of the marginals and coupling 115

10 Application: MCMC **118**
10.1 Motivation 118
10.2 How to use posterior samples 119
10.3 Examples of MCMC algorithms on Ising models 120
10.4 Gibbs sampling 120
10.5 Metropolis-Hastings (MH) algorithms 122
10.6 Irreducibility of MCMC algorithms 123

1 Motivations

Why learn probability theory?

1. Probability theory is a fundamental tool in statistics, computer science, physics, econometrics, and many more traditionally quantitative fields. It is also quickly gaining in importance in fields making the quantitative transition, for example biology, linguistics, sociology, and many more.

2. The theory is beautiful in its own right. Probability Theory can also be approached as a branch of pure mathematics.

Both of the above points are excellent motivations. However by the nature of this course (we are in a stats department!), I will focus on the first point above and heavily use practical motivations throughout the notes.

Why is probability a fundamental tool in so many fields? Because Probability Theory is useful for creating *models*.

Models are sketches of reality that capture the essential of a problem while being amenable to mathematical analysis. Probability-based models are great at incorporating phenomena like uncertainty and non-linearity. Moreover, compared to other types of models you might be familiar with (e.g. linear algebra or calculus based), they arguable tend to be more resilient to *mis-specification*, i.e. they can recover from certain mismatches between the model and reality (data) in the sense of still giving good predictions.

There are many other motivations. I will just mention two more quickly:

From model to prediction: in Bayesian statistics, probability theory is essential not only to formulate models, but also to make predictions. Bayesian statisticians construct a probability model in which both the known quantities (data) and the unknown quantities are modelled using random variables. In Bayesian inference, prediction then involves conditioning on the data. We will use Bayesian statistics as a recurrent example when talking about conditioning.

The computational power of randomness. Probability Theory arises in a surprising way in the subfields of computer science concerned with the design and analysis of algorithms. Researchers have found since the 1940s many problems where the best way to solve deterministic problems is to introduce artificial randomness in the execution of the algorithm (an idea called *algorithmic randomness*). Consider for example the problem of quickly approximating the volume of an arbitrary convex body. In a 1991 landmark paper [2], Dyer, Frieze and Kanna devised the first provably efficient approximation algorithm, which crucially depends on algorithmic randomness to perform random walks. Moreover, it is known that no efficient deterministic algorithms can provide accurate approximations [1]. There are many instances where there are no known deterministic algorithms and where algorithmic randomness is necessary to scale to large problems.

2 Foundations

On being formal: An important thing to realize is that despite the fact that Probability Theory is used to model uncertainty, it is as formal as any other fields of mathematics (despite what you may have been led to believe if you took a course using the standard undergraduate way of teaching). Formalization of the field was achieved by Kolmogorov in the 1930s, when he realized that the same tools used to formalize the notions of area and volume (measures) could be applied to probabilities.

On intuition: While being formal is useful, intuition is important too. Probability is very connected to the real world so using your intuition is also very useful in making guesses that you can then prove using theory.

Why measure theory? Measure theory is the standard framework used to formalize probability theory. There are several reasons to learn the basics of measure theory (i.e., as Pollard puts it, to be *user* of measure theory). We will see some of these technical motivations as we go along:

1. Unified treatment of things are taught separately in naive undergraduate probability course: discrete/continuous, univariate/multivariate.

2. Certain results about expectations are easier to state and hold more generally under the measure-theoretic definition of expectations.

3. Establishing independence can in certain case be much easier under the measure theoretic definition of independence.

However in my view the main motivation is that a big chunk of the literature is written in the language of measure theory. This course will prepare you so that you can be fearless when reading the stat literature.

The danger is to be too formal and that the notation gets in the way of the intuition. I will avoid this pitfall. I will also skip the tedious details that I find less useful in statistics, e.g. certain proofs of existence, especially if they do not reveal a technique more broadly applicable

2.1 The basic vocabulary of probability

The axioms of probability, formulated by Kolmogorov in the 1930s, provide a vocabulary and basic set of rules used in everything that follows. We look at how they become alive by showing how they are used to build two simple models (the second, not as simple as it initially looks!).

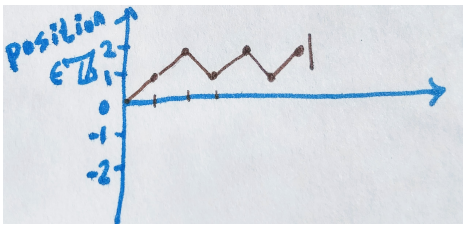
Recurrent example: imagine an infinite railroad, modelled by the integers $\mathbb{Z} = \{\dots, -3, -2, -1, 0, 1, 2, 3, \dots\}$. A train starts at position zero. At every time step,the train operator flips a coin. If the coin shows heads, the train moves left by one unit. Otherwise, the train moves to the right by one unit.

We will look at two versions of this example, version A, where there is fuel limit which restricts the train to six moves, and version B, where there is no fuel limit.

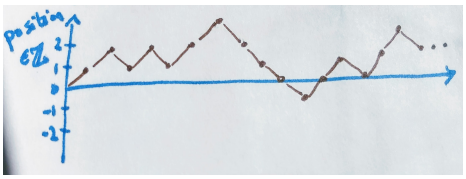
Let us start with version A.

Outcomes/scenarios: several scenarios can arise from the dice-driven train "story" or "experiment". For example the train could go left, right, left, right, right, left. Or it could go left, right, right, right, right, left. We will call these two scenarios, two *outcomes*. The usual letter for one outcome is ω or s.

To help visualization, we will draw one outcome in the time series example as a graph where the x-axis is time t and the y-axis is the position at time t. Here is one example for version A:



and for version B (infinite fuel):

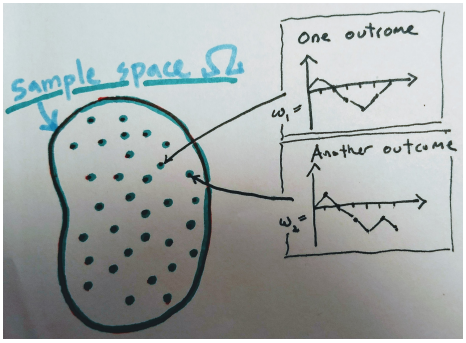


Definition: The set of all possible outcomes is called the *sample space*, which I will denote by Ω or S.

Example: what is the size or cardinality of Ω (i.e. the number of elements in it) for version A of our train example? I will denote the size by $|\Omega|$. Answer:2^6 because there are six bits each free to take two possible values. We conclude that the probability of any given outcome in version A is $1/64$. Note that

$$\sum_{\omega \in \Omega} \text{probability of outcome } \omega = 1.$$

Here is a picture for the train version A example:

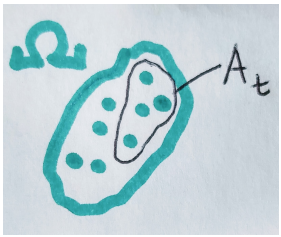


Science-fiction interpretation: if you read or watch science fiction (e.g. check out "The man in the High Castle" by Philip K. Dick), you may think about an outcome as encoding all the information for one universe, and the the set of all outcomes (the sample space) as the *multiverse*.

Note: an outcome may contain many bits of information. We only need one of those bits to get two different outcomes. I.e. outcomes can be very similar, but not identical.

Definition: a set of outcomes is called an *event*.

Example: events are often created using a property defining what outcomes are in the event. For example, consider the set of outcomes where the train goes left in the second time step, A_1. We use the terminology "the event that the train goes left in the second time step." What is its size? Notice that events that are described by a small number of properties tend to be large, while event that are described by a larger number of properties tend to be smaller.

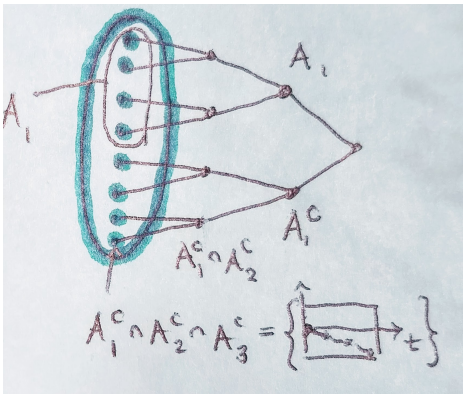


Partitions: often we consider not one event but several events at once. For example we want to categorize outcomes into sub-cases. This can be done using partitions. A *partition* of a set E (for example $E = \Omega$), is a collection[1] of events $B_1, B_2, \ldots$ such that (1) $B_i \cap B_j = \emptyset$ for all $i \neq j$ and (2) $\cup B_i = E$. Each B_i is called a block.

Decision trees are useful to organize a hierarchy of events. Each node in the tree is an event. At the root, we put Ω. When $|\Omega| < \infty$, at the leave we have *singletons*, i.e. sets with only one element. In between, each node of the tree is

[1] Just a synonym for set.

split into subcases. More formally, say we are at a node corresponding to event E. Pick a *partition* $B_1, B_2, \ldots$ is of E. Then the children of E in the decision tree are defined as the blocks of the partition splitting E.



Additivity: it is not too hard to see in example A that the probability of *disjoint* events (i.e. non-intersecting events) can be added up. E.g.:

$$\begin{aligned}&\text{probability go left in first turn} + \text{probability go right in first two turns}\\ &= \text{probability go left in first turn or right in first two turns.}\end{aligned}$$

Both sides are equal to $1/2 + 1/4$.

Note: it is not too hard to see in example A that the probability of *overlapping* events (i.e. intersecting events) cannot always be added up. E.g.:

$$\begin{aligned}&\text{probability go left in first turn} + \text{probability go left in first two turns}\\ &\neq \text{probability go left in first turn or left in first two turns.}\end{aligned}$$

The left hand side is equal to $1/2 + 1/4$ while the right hand side is equal to $1/2$.

A routine task in probability problems consists in expressing events of unknown probability in terms of events of known probability.

Exercise/example: express, in version A of the train example, the event that the train eventually returns home (position zero), in terms of the events:

$$A_t = \text{event that the train goes left in the } t\text{-th time step,}$$

which have known probability. Do it using only the set theoretic operations $\cup$, $\cap$, and set complement $A_t^{\complement}$.

2.2 Motivating the axioms of probability

Difficulty pre-Kolmogorov/pre-1930s: people were already very good at doing computations for things like version A of the train. But surprisingly, they

were running into serious foundational problems when trying to formalize the theory for version B of the problem!

Difficulty: in version B, we cannot always define probability of events by adding up probability of individual outcomes. We cannot have these three things:

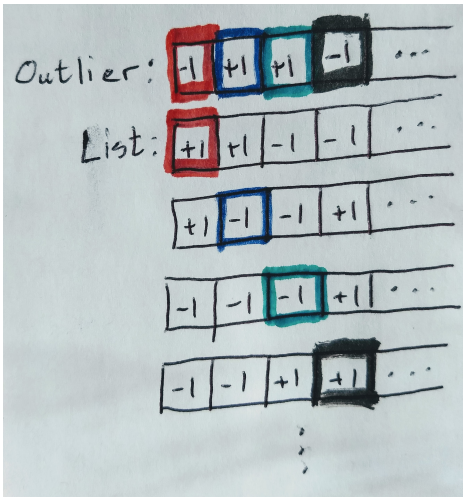
$$\text{probability of one path} = \lim_{t\to\infty}(1/2)^t = 0$$

$$\text{probability of } \Omega = 1$$

$$\sum_{\omega\in\Omega} \text{probability of outcome } \omega = \text{probability of } \Omega.$$

The problem is that each individually appears to make sense (the third one, in the light of the disjoint additivity observation), but they are inconsistent! We will see that the third one is the one causing problems.

Root cause of the problem: in version B, the set Ω is *uncountably infinite*. Recall that a set is countable if we can come up with a list of all the elements in it. More formally, a set is countable if we can come up with a surjective function f taking as input an integer and returning an element in B (think of the input of f as the position in the list, and the output as the item listed in that position). The terminology surjective means $\{f(i) : i \in 0, 1, 2, \dots\} = \Omega$, i.e. the list is exhaustive. To see why the set Ω of all infinite train paths is uncountable, argue as follows: suppose on the contrary that there was a list (i.e. suppose you claim to come up with a surjective function f listing all the paths). I will show you there must be at least one path not in your list, an "outlier." Recall that a path is just a list of coin toss, say encoded as 0 and 1. Here is a picture followed by a description:



1. I start by building the **first** coin toss in my outlier. I look at the first toss of the first element in your list, $f(1)_1$. If I see a 0, I set my first toss to 1. If I see a 1, I set my first toss to 0.

2. Then I build the **second** toss in my outlier. I look at the **second** toss of the **second** element in your list, $f(2)_2$. If I see a 0, I set my second toss to 1. If I see a 1, I set my second toss to 0.

3. **Third** toss: I look at the **third** toss of the **third** element in your list, $f(3)_3$. If I see a 0, I set my third toss to 1. If I see a 1, I set my third toss to 0.

4. etc.

By construction, the outlier cannot be in your list! We conclude the set Ω of all infinite coin tosses is not countable (i.e., *uncountable*).

Remedy: here is the first part of Kolmogorov's insight (in turn based on insights from the then-nascent theory of measure)

1. Give up defining the probability by assigning it to single outcomes.

2. Instead, assign probability to events. I.e. define $\mathbb{P}(A)$ for event A, not $\mathbb{P}(\omega)$ for outcome ω.

3. The function $\mathbb{P}$ is now defined on a much bigger space! Many functions defined on that large space make no sense (e.g. we could have $\mathbb{P}(\Omega) \neq 1$ which does not make sense). Let us use the intuition gained with example version A to extract what are the things we want to assume on $\mathbb{P}$. The list of what to include in order to get an interesting theory is surprisingly short:

   (a) Additivity of disjoint events.

   (b) $\mathbb{P}(\Omega) = 1$.

Second difficulty: we need to be careful about how we define the disjoint additivity axiom! If we allowed additivity over uncountable collections of events, we would be back the "paradox" at the beginning of the section, i.e. that $\mathbb{P}(\{\omega\}) = 0$ but

$$\mathbb{P}(\cup_{\omega\in\Omega}\{\omega\}) = P(\Omega) = 1.$$

Remedy (continued): assume disjoint additivity only for countable collections of events. More formally, assume that if $A_1 \in \mathcal{F}, A_2 \in \mathcal{F}, A_3 \in \mathcal{F}, \ldots$ is a countable collection of disjoint events ($i \neq j \Longrightarrow A_i \cap A_j = \emptyset$), then

$$\mathbb{P}(\cup A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Third difficulty: a strange theorem of measure theory states that there exists no probability distribution defined on all events, $\mathbb{P} : 2^{\Omega} \to [0,1]$ such that $\mathbb{P}(\{\omega\}) = 0$ for all $\omega \in \Omega$.

Intuition: when Ω is uncountably infinite, 2^Ω is very strange! E.g. some events cannot be described by any language (any languages that we know of proceeds by putting letters from a finite alphabet one after another; well, that can only produce a countable set of descriptions. So that mean certain paths in Ω cannot be described in version B of the train example)!

Remedy (continued): do not attempt to define $\mathbb{P}$ on all subsets of Ω. Define it on a subset of events called a σ-algebra. Make the σ-algebra big enough to do all the computations we are interested in doing: taking countable unions, intersections, and complements. As we will see soon this solves our problem, in the sense that there is a $\mathcal{F} \subsetneq 2^\Omega$, $\mathbb{P} : \mathcal{F} \to [0,1]$ such that $\mathbb{P}(\{\omega\}) = 0$ but for example $\mathbb{P}(A_t) = 1/2$.

2.3 The axioms of probability

To summarize the discussion of the last section, a probability space contains three things:

1. a set Ω, called the sample space,
2. a closed collection of events, $\mathcal{F} \subset 2^\Omega$, called a σ-algebra,
3. a probability measure (synonym: probability distribution), $\mathbb{P} : \mathcal{F} \to [0,1]$.

where:

1. the σ-algebra satisfies:
   (a) $\Omega \in \mathcal{F}$,
   (b) $A_1 \in \mathcal{F}, A_2 \in \mathcal{F}, A_3 \in \mathcal{F}, \ldots, \Longrightarrow \cap_{i=1}^\infty A_i \in \mathcal{F}$,
   (c) $A \in \mathcal{F} \Longrightarrow A^\complement \in \mathcal{F}$.
2. and the probability measure satisfies:
   (a) $\mathbb{P}(\Omega) = 1$,
   (b) if $A_1 \in \mathcal{F}, A_2 \in \mathcal{F}, A_3 \in \mathcal{F}, \ldots$ is a countable collection of disjoint events ($i \neq j \Longrightarrow A_i \cap A_j = \emptyset$), then

$$\mathbb{P}(\cup A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Measure: a measure space is a very close cousin to a probability space. The axioms are identical, except for two changes:

1. We remove the bound $[0,1]$ and replace by $[0,\infty)$, and use the terminology measure (often denoted μ) instead of probability, $\mu : \mathcal{F} \to [0,\infty)$.
2. We modify $\mathbb{P}(\Omega) = 1$ into $\mu(\emptyset) = 0$.

2.4 Some basic properties

Some examples of basic properties we can derive from these axioms:

1. $\mathbb{P}(A^{\complement}) = 1 - \mathbb{P}(A)$,
2. if A and B are events that are not necessarily disjoint, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Idea: partition sets into disjoint bits, so that the disjoint additivity can be used. For part 1, use that A and $A^{\complement}$ form a partition of Ω. For part 2, define the set subtraction as $E \backslash F = \{e \in E : e \notin F\}$, and use the partition $A \backslash B, B \backslash A, A \cap B$.

Exercise: write down the argument.

2.5 Probability spaces as models

2.5.1 Coin tosses

Question: you toss two coins, what is the probability of two heads? Note that you are not able to tell the two coins apart.

Sample space: which of these should we pick?

1. $\Omega_1 = \{(H, H), (H, T), (T, H), (T, T)\}$
2. $\Omega_2 = \{\{H, H\}, \{H, T\}, \{T, T\}\}$.

The best answer is Ω_1, but why? Choosing between these two (essentially, selecting one of these two models) is not part of probability theory per se. Use have to use your intuition about the real world here. For example, note that if you painted one coin red and one blue, the setup of this experiment would not have changed. Hence, the model that is most useful is the one that uses lists even though we could not observe this distinction. Probability theory comes in once we have built a model, at which point inference can be carried using mathematical principles. Probability theory can help selecting model though, for example by making certain predictions for a given model, which can then be tested (for example, the long term behavior of frequencies, which is mathematically understood for a wide range of probability models).

2.5.2 Reliability

Reliable systems replicate a critical component (e.g. a power supply in a computer server) so that the whole system works as long as at least one of the two copies works. Consider an assembly line for computer servers. Suppose the first power supply assembly line is observed to put in a power supply that works 60% of the time. The second power supply assembly line is observed to put in a power supply that works 70% of the time. At delivery, both power supplies work 40% of the time.

Exercise: What is the probability that both power supplies are broken at delivery? Hint: the answer is not 12%.

If you answered 12%, you are using a property that is *not* built into (or derivable from) the axioms of probability: namely that if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. It is an extra assumption called *independence* of the events A and B. It is kept separate because there are situation where it is useful to describe the world, and others where it is not (can you imagine a scenario where the two assembly lines are not independent?). This contrast with the disjoint union axiom of probability, which models a universal aspect of reality.

2.6 Statistical models

Statistical models are built using probability models. Consider the following example:

Estimation: consider the train example, but where instead of a standard coin being used to make the decision at each step (50%-50% to go left or right), a *biased* coin is used, i.e. where the train goes left with probability $\theta \in [0, 1]$ and right with probability $1 - \theta$. The problem is that we do not know θ! Instead, we try to reconstruct θ from data (observed paths). This problem, point estimation, is one an important type of problems considered in statistics.

Frequentist models: use not one but many probability distributions all defined on a shared space $(\Omega, \mathcal{F})$:

$$\text{frequentist model} = \{\mathbb{P}_\theta : \mathcal{F} \to [0, 1], \theta \in \Theta\}.$$

If the index set Θ is some subset of $\mathbb{R}^d$, the model is called parametric, otherwise, it is called non-parametric.

Bayesian models: use only one probability distribution. Augment the space Ω to include the unknown quantity $\theta \in [0, 1]$, i.e. $\bar{\Omega} = [0, 1] \times \Omega$. We will go over this way of modelling in more detail when we talk about conditioning.

2.7 Simple examples of σ-algebra

We have seen one "foundational" motivation for σ-algebras. We will also see a more useful motivation soon which is that σ-algebras can encode rules of a game, but we will need to define random variables first. For now, let us look at some examples to make this concept a bit more concrete.

Power set: the power set 2^Ω is always a σ-algebras on Ω. For example

$$\mathcal{F}_0 \coloneqq \{\{a, b, c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a\}, \{b\}, \{c\}, \emptyset\}$$

is a σ-algebras on $\Omega \coloneqq \{a, b, c\}$.

Another discrete example: the collection of events

$$\mathcal{F}_1 \coloneqq \{\{a, b, c\}, \{a, b\}, \{c\}, \emptyset\}$$

is also a σ-algebra on $\Omega \coloneqq \{a, b, c\}$.

A non-example: the collection $\mathcal{F}_2 \coloneqq \mathcal{F}_1 \cup \{\{a\}\}$ is not a σ-algebra. Why? We have $\{a\} \in \mathcal{F}_2$ yet $\{a\}^{\complement} \notin \mathcal{F}_2$.

2.8 More interesting examples of σ-algebra via generation

Let $\mathcal{S}$ denote a collection of events. The machinery of this section is useful when the collection $\mathcal{S}$ is "broken," i.e. when it is not a σ-algebra (example: $\mathcal{S} \coloneqq \mathcal{F}_2$ above).

We would like to "repair" $\mathcal{S}$ by adding more events until we get a closed collection of event. How can we do this is such a way that the output of the repair process is unique and well-defined?

1. Let $\mathcal{F}$ and $\mathcal{F}'$ be two σ-algebra on Ω. **Exercise:** convince yourself that their intersection $\mathcal{F} \cap \mathcal{F}'$ is also a σ-algebra.

2. The result in 1 can be generalized: if we have any collection of σ-algebras $\{\mathcal{F}_\alpha : \alpha \in I\}$, where I is some index set (not necessarily countable), then

$$\bigcap_{\alpha \in I} \mathcal{F}_\alpha$$

   is also a σ-algebra.

3. Let us pick

$$\{\mathcal{F}_\alpha : \alpha \in I\} = \{\mathcal{F}_\alpha : \mathcal{F}_\alpha \text{ is a } \sigma\text{-algebra containing } \mathcal{S}\},$$

   then we get that the intersection of the $\mathcal{F}_\alpha$ is a σ-algebra.

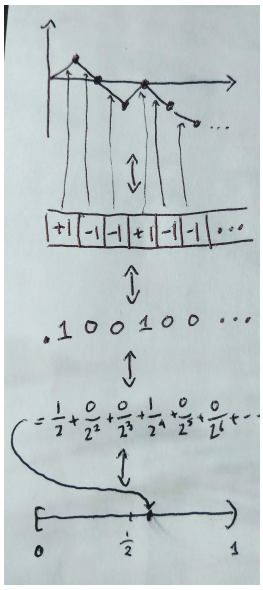
4. We call this intersection the σ-algebra *generated* by $\mathcal{S}$, denoted $\sigma(\mathcal{S})$.

5. Another way to think about $\sigma(\mathcal{S})$: the *smallest σ-algebra containing $\mathcal{S}$*.

Some examples:

1. $\sigma(\mathcal{F}_2) = 2^{\{a,b,c\}}$.

2. An important example where the generated σ-algebra is smaller than 2^Ω: *the Borel σ-algebra*. To follow this example, it will help to first make the observation that infinite paths can be made in correspondence with the interval $[0, 1)$. This is done via the *binary representation* of real numbers, which works schematically as follows:



This correspondence shows that many events of interest can be expressed as intervals in $[0, 1)$. For example what is the interval corresponding to "the train goes left in the first step"? The interval $[0, 1/2)$. So to be able to have these intervals in our σ-algebra, we proceed as follows:

(a) Take $\Omega \coloneqq [0, 1)$, and $\mathcal{S}_{\mathrm{B}} \coloneqq \{F : F \text{ is a finite collection of intervals}\}$.

(b) Convince yourself that $\mathcal{S}$ is not a σ-algebra.

(c) We call $\sigma(\mathcal{S}_{\mathrm{B}})$ the *Borel σ-algebra*, denoted $\mathcal{F}_{\mathrm{B}}$.

(d) Elements of $\mathcal{F}_{\mathrm{B}}$ are called Borel events.

(e) Some measure theory shows that $\mathcal{F}_{\mathrm{B}}$ is a proper subset of $2^{[0,1)}$.

2.9 Simple examples of probability measures

Discrete probability measure: assume you are given a countable sample space Ω and a function $p : \Omega \to [0, 1]$ called a Probability mass function (PMF). Define, for any event $A \in \mathcal{F}$:

$$\mathbb{P}(A) \coloneqq \sum_{\omega \in A} p(\omega).$$

Check that this definition satisfies the axioms of probability. Note: do not confuse p and $\mathbb{P}$, as they take different types of inputs!

Conditional probabilities: fix an event E which you can interpret as some data, with $\mathbb{P}(E) > 0$. Define a new probability distribution as

$$\mathbb{P}'(A) = \frac{P(A \cap E)}{\mathbb{P}(E)}.$$

Check $\mathbb{P}'$ is indeed a probability. If $\mathbb{P}$ is some agent's belief before observing E, $\mathbb{P}'$ can be interpreted as the optimal way for the agent to update its belief after observing E. This updated belief is notated $\mathbb{P}'(\cdot) = \mathbb{P}(\cdot|E)$ and is called the conditional probability given E. From now on, if we write $\mathbb{P}(A|E)$ we will assume $\mathbb{P}(E) > 0$ implicitly.

The uniform probability measure: some (surprisingly heavy) measure theory shows that there exists a probability measure on $\mathcal{F}_{\mathrm{B}}$ such that $\mathbb{P}([a, b)) = b - a$ for all $0 \leq a \leq b < 1$.

This second example is actually enough to build a rich theory! How? Using random variables and their distributions, which we will cover shortly.

Example of a measure: the Lebesgue measure is the same as the uniform probability distribution, except that it is defined on $\mathbb{R}$ instead of $[0, 1)$.

2.10 Computations for discrete models

We already have enough tools to formulate and solve many interesting problems involving *discrete* models (i.e. with $|\Omega| < \infty$).

Exercises: You can solve the following problems by explicitly enumerating all the possible scenarios. However, I ask here that you go beyond and that you formalize your reasoning, by (1) setting up a probability model using set theory only, no English allowed,[2] and (2), use the axioms of probability to solve the problem. All you need is the following two tools.

Tool 1: chain rule. For any event A, B,

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B|A) \ \ (= \mathbb{P}(B)\mathbb{P}(A|B)).$$

Tool 2: law of total probability (version 1). Suppose $B_1, B_2, \ldots$ is a partition of Ω, then

$$\mathbb{P}(E) = \sum_i \mathbb{P}(E \cap B_i) \quad \left(= \sum_i \mathbb{P}(B_i)\mathbb{P}(E|B_i) \right).$$

Proposed strategy: build the model and list the known (conditional) probabilities. Write the (conditional) probability you wish to compute. Then reintroduce the "missing" variable in this last probability using the law of total probability and the chain rule.

1. Prove that the two "tools" follow from the axioms of probability.
2. An HIV test has the following two modes of failure:

[2] Obviously, in general I do recommend a mix of English and math for clarity, but probability beginners tend to rely too much on the former.

- When a patient has the disease, the test will still be negative 2% of the time (false negative)
- When a patient does not have the disease, the test will turn positive 1% of the time (false positive)

Given that the test is positive, what is the updated (posterior) probability that the patient is indeed affected by HIV?

Suppose now 25% of the population has HIV. Given that the test is positive, what is the updated (posterior) probability that the patient is indeed affected by HIV? What happens if the prevalence of the disease is very small instead?

3. There are 3 kindergarten classrooms and 9 kindergarten students in a school. The children are lined up and assigned to classrooms in turn. The principal claims the assignment is uniform over all possible assignments. The classrooms have the following capacities:

   - Classroom (a): 4 students
   - Classroom (b): 3 students
   - Classroom (c): 2 students

   You are fifth in the list of children. What is the probability that you get assigned to class (a)?

2.11 Exercise set 1

1. Solve the exercise in Section 2.4
2. Solve the exercise in Section 2.5.2
3. Solve the exercise in Section 2.10

2.12 Solutions for exercise set 1

2.12.1 Basic properties

First, not that additivity for a countable collection implies additivity for a pair, by taking $A_3 = A_4 = A_5 = \emptyset$. Since A and $A^{\complement}$ are disjoint, $P(A) + P(A^{\complement}) = 1$, and hence the first simple property.

For the second property we make use of the following:

1. $\{A\backslash B, A \cap B\}$ is a partition of A; hence $\mathbb{P}(A\backslash B) + \mathbb{P}(A \cap B) = \mathbb{P}(A)$;
2. $\{B\backslash A, A \cap B\}$ is a partition of B; hence $\mathbb{P}(B\backslash A) + \mathbb{P}(A \cap B) = \mathbb{P}(B)$;
3. $\{A\backslash B, B\backslash A, A \cap B\}$ is a partition of $A \cup B$; hence $\mathbb{P}(A\backslash B) + \mathbb{P}(B\backslash A) + \mathbb{P}(A \cap B) = \mathbb{P}(A \cup B)$.

Now substitute $\mathbb{P}(A\backslash B)$ in the equation in 3 above using equation in 1 above as well as $\mathbb{P}(B\backslash A)$ using equation in 2.

2.12.2 Reliability problem

Define $\Omega = \{(0,0),(0,1),(1,0),(1,1)\}$ where element $\omega = (i,j)$ encodes if the first power supply works ($i = 1$, otherwise $i = 0$) and if the second power supply works ($j = 1$, otherwise $j = 0$). Define W_k as the event that power supply $k \in \{1,2\}$ works, i.e. $W_1 = \{(1,0),(1,1)\}$ and $W_2 = \{(0,1),(1,1)\}$. We know:

1. $\mathbb{P}(W_1 \cap W_2) = 4/10$, (by the way sometimes denoted just $\mathbb{P}(W_1 W_2)$),
2. $\mathbb{P}(W_1) = 6/10$,
3. $\mathbb{P}(W_2) = 7/10$.

The goal is to compute $\mathbb{P}(W_1^{\complement} W_2^{\complement})$.

First, get rid of the complement. How to do so? Use a result from set theory:

De Morgan's "Laws": distributing complements swap unions to intersections and vice versa, i.e. $(A \cup B)^{\complement} = A^{\complement} \cap B^{\complement}$; $(A \cap B)^{\complement} = A^{\complement} \cup B^{\complement}$. E.g., using an example from wikipedia, the search query "NOT (cars OR trucks)" is the same as "(NOT cars) AND (NOT trucks)".

Then we have using De Morgan and the first property from previous question:

$$\mathbb{P}(W_1^{\complement} \cap W_2^{\complement}) = \mathbb{P}((W_1 \cup W_2)^{\complement}) = 1 - \mathbb{P}(W_1 \cup W_2).$$

Finally, using the second property:

$$\mathbb{P}(W_1 \cup W_2) = \mathbb{P}(W_1) + \mathbb{P}(W_2) - \mathbb{P}(W_1 \cap W_2) = 6/10 + 7/10 - 4/10 = 9/10.$$

Hence the answer is $1 - 9/10 = 1/10$.

2.12.3 Discrete computation

Tools: Chain rule follows directly from the definitions:

$$\mathbb{P}(A)\mathbb{P}(B|A) = \mathbb{P}(A)\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \mathbb{P}(A \cap B).$$

To establish the law of total probability (version 1), first show that $\{E \cap B_1, E \cap B_2, \ldots, \ldots\}$ is a partition of E. The result then follows from countable additivity.

Bayes rule: For the HIV test, define the following events: H, the event that HIV is present, and E the observation (a test turned positive). We know $\mathbb{P}(H) = 25\% = 0.25$, $\mathbb{P}(E|H^{\complement}) = 1\%$ and $\mathbb{P}(E^{\complement}|H) = 2\%$. To help visualize this, you may want to draw a decision tree where the first branching is done according to H and $H^{\complement}$, and the second branching, by further refining via intersections with E and $E^{\complement}$. We seek to compute $\mathbb{P}(H|E)$.

Notice that H and $H^{\complement}$ form a partition of Ω. Using this fact and the law of total probability, we obtain what is called Bayes rule:

$$\begin{aligned}\mathbb{P}(H|E) &= \frac{\mathbb{P}(H \cap E)}{\mathbb{P}(E)} \\ &= \frac{\mathbb{P}(H)\mathbb{P}(E|H)}{\mathbb{P}(H)\mathbb{P}(E|H) + \mathbb{P}(H^{\complement})\mathbb{P}(E|H^{\complement})} \\ &= \frac{0.25 \cdot (1 - 0.02)}{0.25 \cdot (1 - 0.02) + 0.75 \cdot 0.01} \approx 97\%\end{aligned}$$

Now what happens if the prevalence ρ of the disease is very small? Generalizing the equation above, we get:

$$\mathbb{P}_{\rho}(H|E) = \frac{1}{1 + \frac{1-\rho}{\rho}\frac{0.01}{1-0.02}},$$

hence as $\rho \downarrow 0$, we get $\mathbb{P}_{\rho}(H|E) \to 0$, which is kind of surprising.

Combinatorics problem:

$\Omega = \{(C_1, C_2, C_3) : \{C_1, C_2, C_3\}$ partitions $\{1, 2, \ldots, 9\}, |C_1| = 4, |C_2| = 3, |C_3| = 2\}$.

The definition of uniform assignment is then, for any $S \subset \Omega$:

$$\mathbb{P}(S) = \frac{|S|}{|\Omega|}.$$

Here we are interested in an even T which can be described as:

$T = \{(C_1, C_2, C_3) : \{C_1, C_2, C_3\}$ partitions $\{1, 2, \ldots, 8\}, |C_1| = 3, |C_2| = 3, |C_3| = 2\}$.

A first solution is to compute $|\Omega|$ and $|T|$. To do so use a decision tree with in the first level, events capturing the assignment of the first classroom:

$$E_C^1 = \{\omega \in \Omega : \omega_1 = C\},$$

then, at the next two levels, events capturing the assignment of the second and third classroom:

$$E_C^i = \{\omega \in \Omega : \omega_i = C\}.$$

Using the fact that the number of subsets of size k from an inventory of size n is $\binom{n}{k}$, you will find that Ω can be decomposed as a regular tree with branching factors $\binom{9}{4}$, $\binom{5}{3}$, 1, and thus:

$$|\Omega| = \frac{9!}{4!3!2!}.$$

Similarly for T,

$$|T| = \frac{8!}{3!3!2!},$$

thus $\mathbb{P}(T) = 4/9$.

A second solution:

$$\Omega = \{(a_1, a_2, \ldots, a_9) : \cup\{a_i\} = \{1, 2, \ldots, 9\}\},$$

where a_i is the assignment of student i, identifying $\{1, 2, 3, 4\}$ as the seats for the first classroom, $\{5, 6, 7\}$, for the second. Then use the events $E_i = \{\omega : \omega_5 = i\}$. We have $|E_i| = |E_j|$ and the event of interest is $E_1 \cup E_2 \cup E_3 \cup E_4$. Thus:

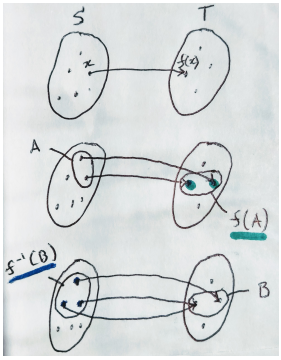
$$\mathbb{P}(E_1 \cup E_2 \cup E_3 \cup E_4) = \frac{4|E_1|}{9|E_1|} = 4/9.$$

2.13 Random variables

Informal definition:

- A random variable is often used to encode a measurement, for example, in the previous train example, whether the train goes left or right at time step 3.
- Random variables are used to model partial observability of the world. In contrast to each outcome $\omega \in \Omega$, which contains all the information possible within the model, a random variable can be defined to "forget" information. For example, you may only known the position of the train at time steps 1 and 6, and what happens in between is unknown.
- Another use of random variables is to express a quantity that is unknown, but that we would like to have the probability of. A concept that I call a "query."

Pre-requisite for the formal definition: if $f : S \to T$ is some function, we can lift function evaluation, which takes as input points and returns points $f(x) \in T$ for $x \in S$, into function evaluation taking as input a set and returning a set, i.e. for $A \subset S$, $f(A) = \{f(x) : x \in A\}$.

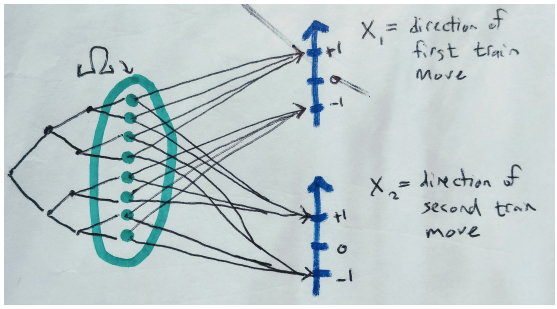


Note that we abuse the notation and use the same symbol for lifted function evaluation, but no confusion can arise because of the type of the input. We do

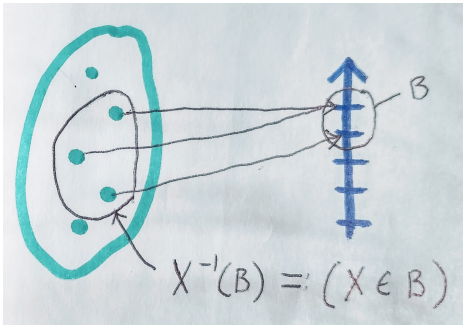
the same for lifted function inverse: for $B \subset T$, $f^{-1}(B) = \{x \in S : f(x) \in B\}$. The lifted inverse is nice because it always exists: the set it returns can be empty or have cardinality larger than one, but it is still a set.

We can now formally define random variables:

1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote a probability space,
2. then a (real) random variable X is:
   (a) a map, $X : \Omega \to \mathbb{R}$,[3]
   (b) such that $X^{-1}(A) \in \mathcal{F}$ for all $A \in \mathcal{F}_{\mathrm{B}}$.



Why do we need condition 2b? Often, we will ask questions like "what states of the worlds can yield an observation in A?" In set theory, the set of such states (outcomes) is $X^{-1}(A)$. Now, we want to be able to compute the probability of this set of outcomes, so we require that $X^{-1}(A)$ be an event (i.e. in the σ-algebra).



Note: random variables are especially interesting when we define more than one on the same sample space.

Synonym: "X is a measurable function."

Notes:

[3]Technically, we add two points to the real line, $+\infty$ and $-\infty$, to ensure for example that limits of say increasing random variables are guaranteed to be random variables even if the sequence diverges for some outcomes ω.

- Observations are not always real numbers, for example, they could be colours, the nodes in a discrete graph, or even a graph. Let $\mathcal{X}$ denotes the set of say all colours that could be potentially observed, $\mathcal{X} = \{\text{blue, red, yellow}\}$. We can modify our definition above to get colour-valued random variables (terminology: "random colours"). This is done as follows:
  - In the definition of real random variable, replace "$\mathbb{R}$" by "$\mathcal{X}$."
  - Since we need a collection of sets on $\mathcal{X}$ in 2b, let us assume we have a (second) σ-algebra $\mathcal{F}_{\mathcal{X}}$ on $\mathcal{X}$ as well as on Ω.
  - Therefore, we see that all we really need is a σ-algebra on the input space Ω, and one on the output space σ-algebra. Notation: a random colour is a $(\mathcal{F} \to \mathcal{F}_{\mathcal{X}})$-measurable function.
- Using this idea, we will later define random vectors (vector-valued random variable), random graphs, random sets, random functions, and even random probability measures!
- (Real) random variables and random vectors are special though, as they will later allow us to define the notion of expectation (not possible with colours, as we cannot for example "add colours"). More on this later!

Terminology: values in the set $\mathcal{X}$ are called *realizations*, and are often denoted with the small-cap version of the random variable, e.g. a realization of X is denoted $x = X(\omega)$. This distinction is useful for example to distinguish $f(x)$, which is just function evaluation, from $f(X)$, which is the construction of a new random variable obtained by composing f and X (composition is reviewed in the next section).

Warning: make sure you follow the type conventions! For example, $\mathbb{P}(\text{red})$ or $\mathbb{P}(X)$ are not defined![4]

2.14 Compositions of random variables

Recall: $g \circ h$ means a new function that first applies h to the input, then plugs in the intermediate quantity into g: $g \circ h(x) = g(h(x))$.

Convention: use capital letters only for the random variables mapping elements from the sample space $\Omega \to \mathcal{X}$. Use standard function notation (g, h, etc) for subsequent transformations $g : \mathcal{X} \to \mathcal{X}'$ of the output of X. E.g. $g \circ X = g(X)$.

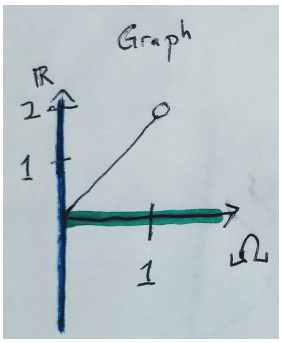
Exercise: show that the composition $g(X)$ of a random variable X with a measurable function g is a random variable.

Convention: from now on, we will implicitly assume all functions involved are measurable.

[4] Some authors do give a meaning to the latter, but it is not what you think! We will see this meaning, namely the expectation of X, later.

2.15 The graph of a random variable

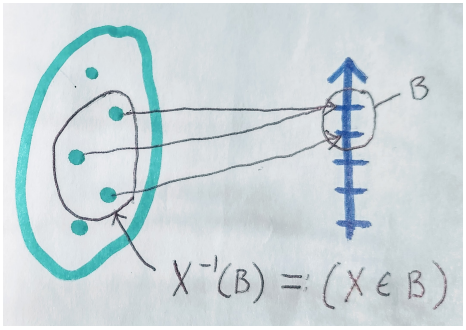
In the special case where $\Omega = \mathbb{R}$, we can plot the *graph* real random variables $X : \Omega \to \mathbb{R}$. While in practice Ω is usually much more complicated, looking at simple examples where $\Omega = \mathbb{R}$ is a powerful technique to get some intuition on several results we will cover in the next few weeks. Note that measurability is much weaker than continuity, so the the graph of random variables can be quite pathological in general.



2.16 The probabilist's event notation

It is tedious to write expressions like $X^{-1}(\{r \in \mathbb{R} : 6 \leq r\})$. Probabilists noticed that the notation $(6 \leq X)$ was not defined, and decided to give it a new, precise meaning: $X^{-1}(\{r \in \mathbb{R} : 6 \leq r\})$. Some other examples:

- $(X \in A) \coloneqq X^{-1}(A)$,
- $(X = x) \coloneqq X^{-1}(\{x\})$.



More generally:

(logical statement s containing a random variable X) $\coloneqq \{\omega \in \Omega : s(X(\omega)) \text{ is true}\}$.

2.17 Constant random variables

The simplest example of a random variable is a constant function, e.g. $X(\omega) = 42$ for all ω or $Y(\omega) = 3.14$ for all ω, denoted $X = 42$ and $Y = 3.14$ respectively. These are boring but important building blocks. We will abuse notation and denote the random variables such as X and Y here by just 42 and 3.14 respectively. For example $Z + 3$ if a shorthand for the composition $g(Z, W)$, where $g(z, w) = z + w$ and $W(\omega) = 3$.

2.18 Indicator random variables

Often we need random variable taking binary values (either zero or one). These are called indicator random variables.

For any set A, define the *indicator function* or just indicator for short as follows:

$$\mathbf{1}_A(\omega) \coloneqq \mathbf{1}[\omega \in A] \coloneqq \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise.} \end{cases}$$

The indicator random variable is just an indicator on an event.

If A can be defined using the probabilist's notation introduced in Section 2.16, for example if $A = (X = x)$, we will use the notation $\mathbf{1}(X = x)$ for $\mathbf{1}_A$.

2.19 Modelling randomized algorithms as random variables

Recall: we mentioned randomized algorithms as one of the motivations for learning Probability Theory. Randomized algorithms can be defined as algorithms having access to the realization of one random variable. Since in practice we need several random variables, randomized algorithms use *pseudo-random generators* to turn one random number into a potentially very large list of random variables. In this section we see how to model randomized algorithms and pseudo-random generators using random variables.

Example of a simple randomized algorithm: which simulates 100 coin flips and counts the number of times the coin comes heads. More precisely, the algorithm takes as input a *random seed* r (an integer, but modelled in a computer as being in a finite range of values, from 1 and say 2^{32}), and proceeds as follows:

1. $s \leftarrow 0$
2. For $i = 1, 2, \ldots, 100$:
   (a) Extract a pseudo-random coin flip from the least significant binary digit of r:
   $$x_i = r \mod 2.$$
   In other words, x_i is equal to one if and only if r is an odd number, otherwise it is equal to zero.

(b) increment the sum counter, $s \leftarrow s + x_i$

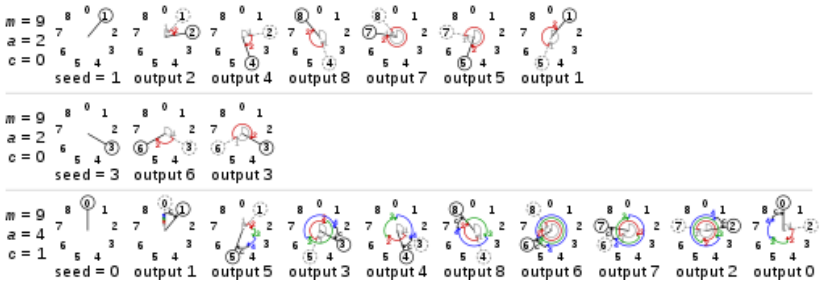
(c) Get a new pseudo-random integer by "flipping" an artificial wheel of fortune (wheel because we do modular arithmetic modulo m, the number of slots in the wheel), first nudging it by a multiplicative amount a, then by a constant amount c:

$$r \leftarrow (ar + c) \mod m.$$

3. Return the sum s

Note: Here, m, a and c are fixed constants. For example, the famous Numerical Recipes book [4] recommends $m = 2^{32}$, $a = 1664525$ and $c = 1013904223$.

Illustration of how the flipping works: from wikipedia, created by user Cmglee, distributed under CC BY-SA 3.0



Model for the algorithm? Take Ω to be the set of integers $1, \ldots, 2^{32}$. A seed corresponds to an outcome $\omega \in \Omega$. Given a seed, everything is deterministic in the algorithm. In particular, we can think about the value x_1 as the output of a deterministic function taking a seed as input. Let us call this function, $X_1 : \Omega \to \{0, 1\}$, and as our choice of letter suggest, it is indeed a random variable. Similarly, we can define random variables $X_2, X_3, \ldots, X_{100}$, each one being a deterministic function (of increasing complexity) of the input random seed ω. E.g.:

$$X_2(\omega) = ((a\omega + c) \mod m) \mod 2.$$

Practical importance. This model illustrates a key design pattern useful when building any randomized algorithm:

- Make the random seed an explicit input of the program.
- Given this input, all computations should be deterministic.

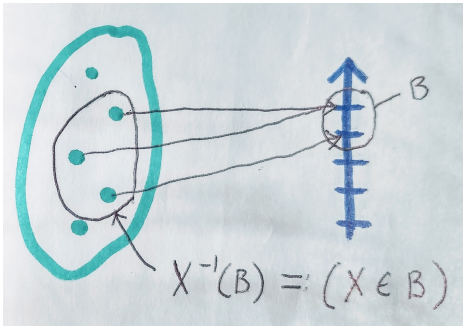
This is useful because you can re-run exactly the same program, i.e. *reproduce your results*. This comes handy if there is a bug you need to fix and you want to replay the crash to analyze in detail what happened. It is also handy if someone wants to replicate somebody else's published work.

Better pseudo-random generators: the recipe used here ($r \leftarrow (ar + c) \bmod m$) is called a *linear congruential generator*. This class of generators have been largely superseded by other methods, a good choice for example is the Mersenne Twister pseudo-random generator [3].

2.20 Distribution of a random variable

Idea: you give me a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable $X : \Omega \to \mathcal{X}$, and I create a new probability $\mathbb{P}'$. This new probability is defined on the values $\mathcal{X}$ that the random variable takes.

Definition: for any $B \in \mathcal{F}_{\mathcal{X}}$, set $\mathbb{P}'(B) \coloneqq \mathbb{P}(X \in B)$.



Notation: we denote this new probability $\mathbb{P}'$ by $\mathbb{P}_X$, and call it the *distribution of* X.

Exercise: suppose X is an indicator variable on an event A, with $p = \mathbb{P}(A)$. Find the distribution of X. The answer is called "Bernoulli with parameter p", denoted $\mathbb{P}_X = \text{Bern}(p)$. The shorthand $X \sim \text{Bern}(p)$ is also widely used in statistics.

Possible confusion: "probability distribution" is a synonym of "probability measure." Here we defined the "distribution *of* X," which is a specific way of constructing a probability measure.

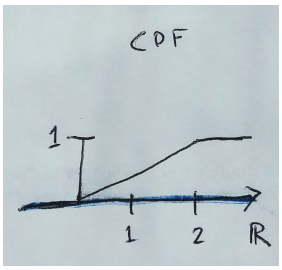
2.21 Cumulative distribution function

Idea: a probability is a function taking inputs from a tricky space ($\mathcal{F}$). This makes it hard to plot naively. It would be nice to be able to summarize it with a function taking inputs in a more familiar space, $\mathbb{R}$.

Note: the following definition only works for $\mathcal{X} = \mathbb{R}$.

Definition: the Cumulative Distribution Function (CDF) of a random variable is given for all $x \in \mathbb{R}$ by:

$$F_X(x) \coloneqq \mathbb{P}(X \leq x).$$



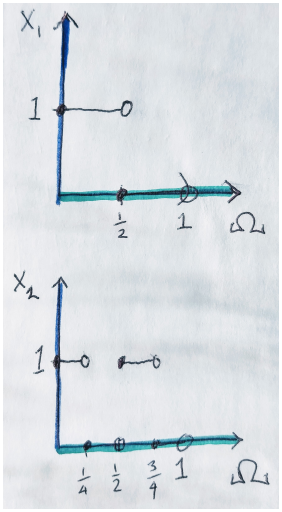
Exercise: this is the same as $\mathbb{P}_X((-\infty, x])$.

2.22 Equality in distribution

Example: Suppose the only probability distribution available in some programming language is the uniform distribution on $[0, 1)$. How can we transform it into a coin toss?

An answer: $X_1 = \mathbf{1}_{[0,1/2)}$.

Question: is this answer unique? No! For example, $X_2 = \mathbf{1}_{[0,1/4)} + \mathbf{1}_{[1/2,3/4)}$ will also do!



Note:

1. $X_1 \neq X_2$ (the two functions are not equal, for example $X_1(1/4) = 1 \neq 0 = X_2(1/4)$
2. but: $\mathbb{P}_{X_1} = \mathbb{P}_{X_2}$.

Definition: We call 2 *equality in distribution*, denoted $X_1 \stackrel{d}{=} X_2$.

2.23 Densities (first definition)

A random variable is said to have *density* $f \geq 0$ if:

$$F_X(x) = \int_{-\infty}^{x} f(z)\,\mathrm{d}z.$$

Note that we will cover a more general definition of density later in this course.

Example: a density for the exponential distribution is given by

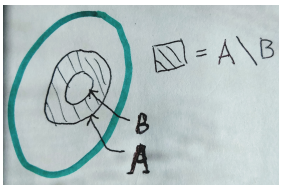
$$f(x) = \mathbf{1}[x \geq 0]\lambda e^{-x\lambda}.$$

Exercise: find an example where there is a x such that a density has $f(x) > 1$.

2.24 Limit properties of probability measures

Lemma: monotonicity. If $B \subset A$ are events, then $\mathbb{P}(B) \leq \mathbb{P}(A)$.

Proof idea: Use the "donut decomposition," $A = B \sqcup (A \backslash B)$, where we use the symbol $\sqcup$ to denote a union while asserting that the two sets we are taking the union over are disjoint.



Notation: to express the following limit properties, we make use of the following overloaded notations,

- Monotone real numbers limits:
  - If $r_1 \leq r_2 \leq \ldots$, and $\lim r_i = r$, we write $r_i \uparrow r$,
  - If $r_1 \geq r_2 \geq \ldots$, and $\lim r_i = r$, we write $r_i \downarrow r$.
- Monotone set limits:
  - If $A_1 \subset A_2 \subset \cdots$, and $\cup A_i = A$, we write $A_i \uparrow A$,
  - If $A_1 \supset A_2 \supset \cdots$, and $\cap A_i = A$, we write $A_i \downarrow A$.

Monotonicity of probability measures:

- $A_i \uparrow A \implies \mathbb{P}(A_i) \uparrow \mathbb{P}(A)$,
- $A_i \downarrow A \implies \mathbb{P}(A_i) \downarrow \mathbb{P}(A)$.

Proof idea for the increasing case: generalize the donut decomposition and write

$$A = A_1 \sqcup (A_2 \backslash A_1) \sqcup (A_3 \backslash A_2) \sqcup \cdots,$$

then use countable additivity to get:

$$\mathbb{P}(A) = \mathbb{P}(A_1) + \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}(A_{i+1} \backslash A_i),$$

using our previous monotonicity property, and telescoping the sum inside the limit, we get:

$$\mathbb{P}(A) = \mathbb{P}(A_1) + \lim_{n\to\infty} [\mathbb{P}(A_n) - \mathbb{P}(A_1)] = \lim_{n\to\infty} \mathbb{P}(A_n).$$

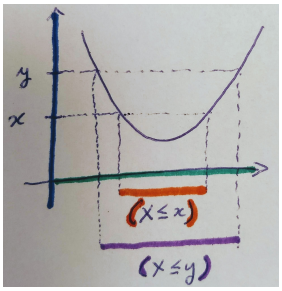
2.25 Limit properties of CDFs

Since the CDF is derived from a probability measure, it shares similar continuity property. But being a function from the real to $[0, 1]$, these monotonicity properties coincide with familiar notions from elementary real analysis.

Notation: Throughout this section, F denotes the CDF of some random variable X, $F := F_X$.

Monotonicity of CDFs: $x \leq y \implies F(x) \leq F(y)$ (F is monotone increasing).

Proof: $x \leq y \implies (X \leq x) \subset (X \leq y)$, so we can use monotonicity of the probability measure $\mathbb{P}$ to conclude the proof.



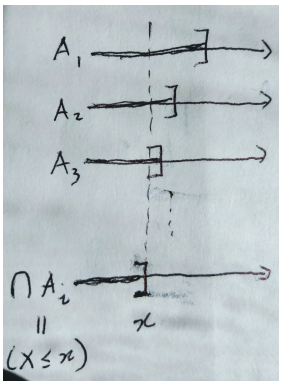
Semi-continuity property:

1. $x_i \uparrow x \implies$ the limit $\lim F(x_i)$ exists,
2. $x_i \downarrow x \implies F(x_i) \downarrow F(x)$.

Proof idea for the decreasing case: we have $(X \leq x_i) \supset (X \leq x_{i+1})$, so by monotonicity of $\mathbb{P}$, we get, for $A_i = (X \leq x_i)$,

$$F(x_i) = \mathbb{P}(A_i) \downarrow \mathbb{P}(\cap A_i),$$

where $\cap A_i = (X \leq x)$ (see figure below), therefore $\mathbb{P}(\cap A_i) = F(x)$.



Reason: for the asymmetry between semi-continuity property 1 and 2. First, do the proof for the increasing case as an exercise. You will see that in the increasing case, the limit of the probabilities is given by $\mathbb{P}(X < x)$, which is not guaranteed in general to be equal to $F(x)$ (because of a potential point mass as x).

Terminology: functions that satisfy the semi-continuity properties (1 and 2) are called cadlag, coming from "continue à droite, limite à gauche" (French for "continuous to the right, limits to the left").

Proposition: let $F : \mathbb{R} \to [0, 1]$ be some function. The following are equivalent:

1. The function F is non-decreasing, cadlag and satisfies the boundary conditions $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to+\infty} F(x) = 1$ (denoted $F(-\infty) = 0$ and $F(+\infty) = 1$),
2. There is some random variable X with $F_X = F$.

Proof idea: we proved the main steps of 1$\Longleftarrow$2. The main idea for the other direction is to use the following construction:

$$X(\omega) \coloneqq \sup\{x : F(x) < \omega\}.$$

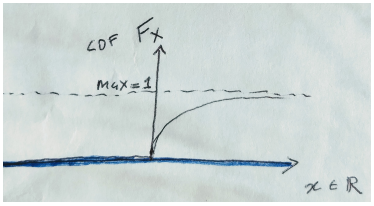
Note: this is a first instance of an important idea in this course, namely to identify the distributions of random variables with simpler types of functions.

2.26 Building random variables with a prescribed CDF

Example/exercise: simulate an exponential random variable (defined below) from a uniform distribution on $[0, 1)$, where an exponential random variable is defined as follows:

Definition: we say X is a standard exponential random variable, a statement denoted $X \sim \text{Exp}(\lambda)$, if

$$F_X(x) = \mathbf{1}[x \ge 0](1 - e^{-x\lambda}).$$



Hint: use the construction from the "proof idea" of Section 2.25, which is called the "inverse CDF."

2.27 σ-algebra, revisited

Often it is useful to put some restrictions on what a random variable can depend on. For example, a model may have two random variables, Y which is observed, and X which is unknown. A statistical estimator δ should only depend on Y, not X.

Method 1: One way to do this is to force δ to be a composition based on Y only: $\delta = f(Y)$ for some f.

Method 2: There is another equivalent way to do that based on the σ-algebra generated by a random variable, defined as:

$$\sigma(Y) = \{(Y \in B) : B \in \mathcal{F}_{\mathrm{B}}\},$$

(Exercise: this is a σ-algebra).

Equivalence of the 2 methods: In an optional question in the assignment, you will show: $\sigma(\delta) \subset \sigma(Y)$ if and only if $\delta = f(Y)$ for some measurable f.

Notation: We use the shorthand $\delta \in \sigma(Y)$ for $\sigma(\delta) \subset \sigma(Y)$.

Exercise: To get a feeling for what this means, consider $X = (X_1, X_2)$ as the value on two dice, but you only observed their sum Y. We want an estimator for X_1 based on the observed sum. Consider $\delta_1 = 6\mathbf{1}[Y > 6] + \mathbf{1}[Y \leq 6]$ and $\delta_2 = X_1$. Check $\delta_1 \in \sigma(Y)$ but $\delta_2 \notin \sigma(Y)$.

Answer:

$$\begin{aligned}
\Omega &= \{(i,j) : i,j \in \{1,2,\ldots,6\} \\
\sigma(X) &= 2^{\Omega} \\
\sigma(Y) &= \sigma(\{\{(1,1)\},\{(2,1),(1,2)\},\{(1,3),(2,2),(3,1)\},\ldots,\{(6,6)\}\}) \\
&= \sigma(\{\{(i,j) : i+j=k\} : k = 2,3,\ldots,12\}) \\
\sigma(\delta_1) &= \sigma(\{\{(i,j) : i+j>6\},\{(i,j) : i+j \leq 6\}\}).
\end{aligned}$$

2.28 Exercise set 2

1. Let X be a random variable with a uniform distribution on $[0,2)$. Draw a possible graph of X, a density for X, and the CDF of X.

2. Solve the exercise in Section 2.8.

3. Solve the exercise in Section 2.14.

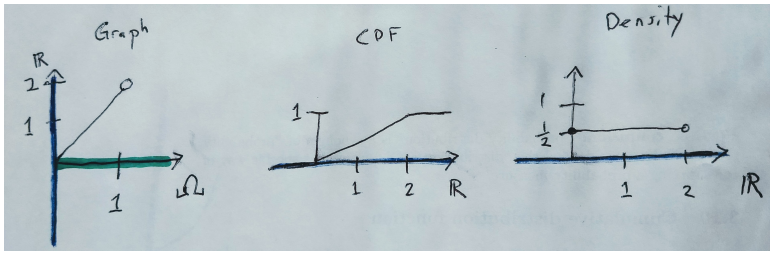
4. Solve the exercise in Section 2.23.

5. Solve the exercise in Section 2.26.

6. If $X_1 \sim F$ (a notation that means that $\mathbb{P}(X_1 \leq x) = F(x)$), and $X_1 \geq 0$, find the CDF of $X_2 \coloneqq X_1^2$.

2.29 Solutions for exercise set 2

1. The three pictures should be as follows:



   (a) Graph of X: The x-axis should be a bounded segment labelled Ω. The y-axis should be the full real line. There are several choices for the function. For example the line $y = x$ or $y = 1 - x$ on the interval $[0, 1)$ are acceptable. Other choices are possible.

   (b) CDF: The x-axis should be the full real line. The y axis should be the interval $[0, 1]$. The function should be zero in $(-\infty, 0]$, then affine in $[0, 1]$, then one in $[1, +\infty)$.

   (c) Density: the x axis should be the full real line. The y axis should be the positive real line. The function should be the indicator on the set $[0, 1)$.

2. We need to check the three conditions given in the definition of σ-algebra:

   (a) Since $\mathcal{F}$ is a σ-algebra, $\Omega \in \mathcal{F}$ and since $\mathcal{F}'$ is a σ-algebra, $\Omega \in \mathcal{F}'$, therefore $\Omega \in \mathcal{F} \cap \mathcal{F}'$.

   (b) We need to show that if $A_1, A_2, \ldots$ are all in $\mathcal{F} \cap \mathcal{F}'$, then $\cap A_i \in \mathcal{F} \cap \mathcal{F}'$. By the definition of intersection, we have that $A_1, A_2, \ldots$ are all in $\mathcal{F}$. Since $\mathcal{F}$ is a σ-algebra, it follows that $\cap A_i \in \mathcal{F}$. By the same reasoning, $\cap A_i \in \mathcal{F}'$. Therefore, $\cap A_i \in \mathcal{F} \cap \mathcal{F}'$.

   (c) We need to show that if $A \in \mathcal{F} \cap \mathcal{F}'$, then $A^{\complement} \in \mathcal{F} \cap \mathcal{F}'$. By the definition of intersection, we have that $A \in \mathcal{F}$. Since $\mathcal{F}$ is a σ-algebra, it follows that $A^{\complement} \in \mathcal{F}$. By the same reasoning, $A^{\complement} \in \mathcal{F}'$. Therefore, $A^{\complement} \in \mathcal{F} \cap \mathcal{F}'$.

3. We have that $X : \Omega \to \mathcal{X}$ and $g : \mathcal{X} \to \mathcal{X}'$ are random variables. Let us denote by $\mathcal{F}_\Omega, \mathcal{F}_\mathcal{X}$ and $\mathcal{F}_{\mathcal{X}'}$ the σ-algebra on Ω, $\mathcal{X}$ and $\mathcal{X}'$ respectively. Let $A \in \mathcal{F}_{\mathcal{X}'}$. We have to show that $(g \circ X)^{-1}(A) \in \mathcal{F}_\Omega$. First, note that $(g \circ X)^{-1}(A) = X^{-1}(g^{-1}(A))$. Since g is a random variable, $g^{-1}(A) \in \mathcal{F}_\mathcal{X}$. Next, since X is a random variables, $g^{-1}(A) \in \mathcal{F}_\mathcal{X}$ implies that $X^{-1}(g^{-1}(A)) \in \mathcal{F}_\Omega$.

4. Examples are common. The uniform distribution on $[0, 1/2)$ has height 2.

5. As in Section 2.25, we pick $X(\omega) \coloneqq \sup\{x : F(x) < \omega\}$ (note that in this special case were $\Omega = [0, 1)$, writing "$F(x) < \omega$" is well defined—for general Ω is would not). This construction has some nice properties (draw a picture of a CDF having flat regions as well as discontinuities to convince yourself, as these are the interesting "corner cases"):

   (a) X is monotone increasing,
   (b) $X \circ F(x) \leq x$,
   (c) $F \circ X(\omega) \geq \omega$.

   We first use properties 5b and 5c to prove the following set equality:

   $$(X \leq x) = \{\omega \in \Omega : \omega \leq F(x)\}.$$

   We show that the LHS includes the RHS and vice versa:

   - $\{\omega \in \Omega : \omega \leq F(x)\} \subseteq (X \leq x)$:

     $$\begin{aligned} \omega \leq F(x) &\implies X(\omega) \leq X \circ F(x) \text{ (from 5a)} \\ &\implies X(\omega) \leq x \text{ (from 5b).} \end{aligned}$$

   - $(X \leq x) \subseteq \{\omega \in \Omega : \omega \leq F(x)\}$:

     $$\begin{aligned} X(\omega) \leq x &\implies F \circ X(\omega) \leq F(x) \text{ (monotonicity of CDFs)} \\ &\implies \omega \leq F(x) \text{ (from 5c).} \end{aligned}$$

   Finally, it follows that:

   $$\begin{aligned} \text{CDF of } X &\coloneqq \mathbb{P}(X \leq x) \\ &= \mathbb{P}\{\omega \in \Omega : \omega \leq F(x)\} \\ &= F(x) \text{ (by the definition of uniform probability).} \end{aligned}$$

   Here in this special case: $X(\omega) = -\lambda^{-1} \log(1 - \omega)$.

6. We have:

   $$\begin{aligned} F_{X_2}(x) &\coloneqq \mathbb{P}(X_2 \leq x) \\ &= \mathbb{P}(X_1^2 \leq x) \\ &= \mathbb{P}(-\sqrt{x} \leq X_1 \leq \sqrt{x}) \\ &= \mathbb{P}(X_1 \leq \sqrt{x}) \text{ (Non-negativity assumption)} \\ &= F(\sqrt{x}). \end{aligned}$$

3 Integration and expectation

3.1 Overview

How to define the mean? At an undergraduate level, this is usually done as follows for continuous random variables:

$$\mathbb{E}[X] \coloneqq \int_{-\infty}^{+\infty} x f(x) \, \mathrm{d}x,$$

where f is the density of X. There are two limitations with this definition:

1. it is not very intuitive (why multiply the density with an x?),
2. we need a separate definition for discrete random variable (and what about cases where we have both continuous and discrete parts?).

Better definition: the expectation is the area under the graph of X!

Note: we will need to generalize the notion of "area," to cover cases where $\Omega \neq \mathbb{R}$.

Terminology: the definition of integral we will cover today is called the Lebesgue integral (not to be confused with the Lebesgue measure). It is the default definition in measure theory, and in general in probability theory.

Note: we will get the undergraduate definition of expectation of a continuous random variable as a special case arising when X has a density. However, the Lebesgue expectation does not need to assume existence of a density.

3.2 Notation, inputs and outputs

The integral you know from calculus (called the Riemann integral) takes one input (a function $f : \mathbb{R} \to \mathbb{R}$) and return one real number. In contrast, the Lebesgue integral needs *two* inputs:

1. a probability measure $\mathbb{P} : \mathcal{F} \to [0, 1]$, where $\mathcal{F}$ is a σ-algebra on a sample space Ω,
2. a random variable $X : \Omega \to \mathbb{R}$.

Notations from real analysis: you will see different notations depending on the author/community to encode this operator on two inputs, for example (these are all synonyms):

- $\int X \, \mathrm{d}\mathbb{P}$
- $\int X(\omega) \mathbb{P}(\, \mathrm{d}\omega)$
- $\mathbb{P}X$

- $(\mathbb{P}, X)$.

In probability theory and Bayesian statistics, there is often a "global" probability $\mathbb{P}$. When this is the case:

Notation in probability theory/Bayesian statistics:

$$\mathbb{E}[X] \coloneqq \int X \, \mathrm{d}\mathbb{P}.$$

Frequentist statistics: in this branch of statistics, there is often a collection of probabilities indexed by a parameter θ, i.e. $\{\mathbb{P}_\theta : \theta \in \Theta\}$. When this is the case:

$$\mathbb{E}_\theta[X] \coloneqq \int X \, \mathrm{d}\mathbb{P}_\theta.$$

3.3 Generalizing the notion of the "area" of a "rectangle"

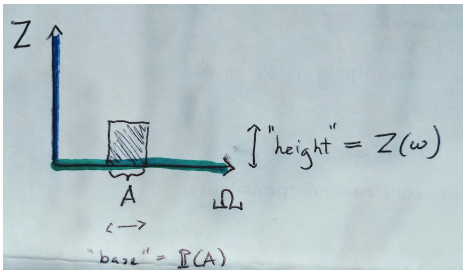
Let us start by defining the notion of area under the curve for something simple: an indicator function multiplied by a constant, $Z = a\mathbf{1}_A$. To make it easier to visualize, let us make this assumption from now on (we will relax it at some point later):

Assumption: assume that all random variables take values ≥ 0.

Now if A is an interval, the graph is just a rectangle! The height is a. What should be the base? Since we are given a probability, let us use it to measure the base, giving $\mathbb{P}(A)$ for the base. This suggests:

Definition: the Lebesgue integral for indicator function is given by

$$\mathbb{E}[Z] = \int Z \, \mathrm{d}\mathbb{P} \coloneqq a\mathbb{P}(A).$$



Note: A could actually be complicated (e.g., the Cantor set), but this definition still holds as long as $A \in \mathcal{F}$.

Note: there will be cases (especially when we talk about limits) where the base has measure zero, but the height is infinite. We would like our definition to return zero in these cases (since the function blows up on a negligeable set). For this reason, we define $0 \times \infty = 0$.

Exercise: compute the expectation of a constant.

Important special case: following from the definition directly, the expectation of an indicator on an event A is just the probability of A, i.e.

$$\mathbb{E}[\mathbf{1}_A] = \mathbb{P}(A).$$

In light of the answer of the exercise in Section 2.20, we only need the distribution of the indicator random variable to compute its expectation. As we will see soon, this is always true, i.e. we only need the distribution of a random variable in order to compute its expectation. In other words, all random variables sharing the same distribution have the same expectation (and many distinct random variables do share the same distribution, see Section 2.22).

Exercise: based on the above special case, can you guess the general formula for computing the expectation of a random variable given only the distribution of the random variable, not the random variable itself?

3.4 Area under the graph of "simple functions"

Simple function: a simple function is a random variable of the form

$$Y = \sum_{i=1}^{N} a_i \mathbf{1}_{A_i},$$

where the A_i are assumed to be disjoint.

Definition: motivated by linearity we extend the definition of the Lebesgue integral to simple functions:

$$\mathbb{E}[Y] = \int Y \, \mathrm{d}\mathbb{P} \coloneqq \sum_{i=1}^{N} a_i \mathbb{P}(A_i).$$

Note: our previous definition for indicators is just a special case of this, so we are not contradicting ourselves.

Property: the above definition is motivated by linearity, and important property we want expectations to enjoy:

$$\mathbb{E}[Y + Y'] = \mathbb{E}[Y] + \mathbb{E}[Y'].$$

Exercise: prove this property holds for the above definition of integral of simple functions.

3.5 Area under graph of non-negative random variables

Now, how to define the area under the graph of an arbitrary non-negative random variable? We make use of our previous definition for simple functions:

Definition: let $X \geq 0$ (meaning $X(\omega) \geq 0$ for all $\omega \in \Omega$),

$$\mathbb{E}[X] = \int X \, \mathrm{d}\mathbb{P} \coloneqq \sup \left\{ \int Y \, \mathrm{d}\mathbb{P} : Y \text{ is simple and } 0 \leq Y \leq X \right\}.$$

Exercises: show $\mathbb{E}[X] \geq 0$ and monotonicity: $X \leq Y \implies \mathbb{E}[X] \leq \mathbb{E}[Y]$.

3.6 Algorithmic construction

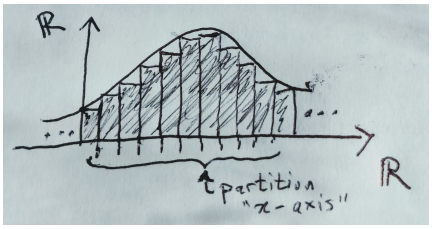
To avoid having to deal with an uncountable collection we follow a two steps strategy:

1. express the random variable to integrate X as the limit of a sequence of increasing and simple random variables. This is done using:
   Proposition: (approximation by simple functions) For any random variable $X \geq 0$, there exists a sequence of random variables $0 \leq Y_1 \leq Y_2 \leq \dots$ such that:
   (a) each Y_n is simple, and
   (b) for all $\omega \in \Omega$, $\lim_{n\to\infty} Y_n(\omega) = X(\omega)$ (this property is known as **pointwise convergence**, denoted $Y_n \to X$, or in this case since the r.v. are additionally increasing, $Y_n \uparrow X$).

2. We will use the **monotone convergence theorem (MCT)** to exchange the limit and integral: if $0 \leq Y_1 \leq Y_2 \leq \dots$ are non-negative random variables (not necessarily simple, although they are in this specific context), then
$$\underbrace{\int (\lim_{n\to\infty} Y_n) \, \mathrm{d}\mathbb{P}}_{\text{hard!}} = \lim_{n\to\infty} \underbrace{\int Y_n \, \mathrm{d}\mathbb{P}}_{\text{easier!}}.$$

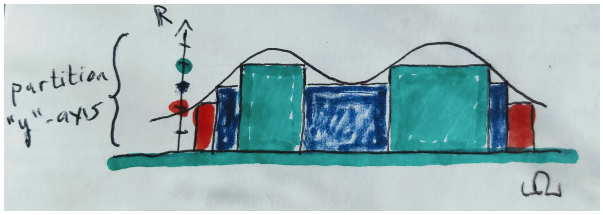
Proof idea of the proposition on approximation by simple functions:

1. Recall that in the case of a Riemann integral, we do something similar, i.e. breaking the x-axis of the graph of X into a grid and making this grid finer and finer.



2. In general, this cannot work here, because the x-axis, Ω, is not necessarily $\mathbb{R}$.

3. Idea: break the y-axis instead! Then use the inverse of the random variable X^{-1}, to get the A_i's required in the definition of simple functions.



3.7 Proving tool: simple function approximation + MCT

The previous section sets the stage for a powerful proving strategy:

1. Suppose you want to prove an identity involving expectations.
   Example: linearity for non-negative random variables $X, X' \geq 0$,

$$\mathbb{E}[X + X'] = \mathbb{E}[X] + \mathbb{E}[X'].$$

2. First prove that the identity holds for simple random variables
   Example: that was an earlier exercise in the case of linearity.

3. Then, use the approximation theorem to get simple $Y_n \uparrow X$ and $Y'_n \uparrow X'$.

4. Use MCT to conclude.
   Example:

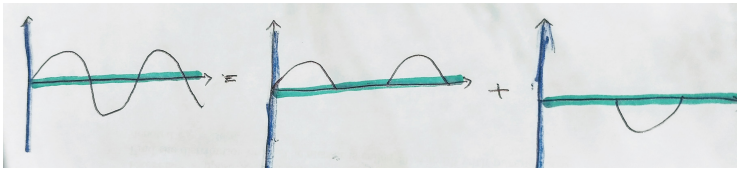
$$\begin{aligned}
\mathbb{E}[X + X'] &= \mathbb{E}[\lim(Y_n + Y'_n)] \\
&= \lim \mathbb{E}[Y_n + Y'_n] \quad \text{(we can use MCT here since } 0 \leq Y_n + Y'_n \uparrow X + X'\text{)} \\
&= \lim(\mathbb{E}[Y_n] + \mathbb{E}[Y'_n]) \quad \text{(easy to prove since } Y_n, Y'_n \text{ are simple)} \\
&= \lim \mathbb{E}[Y_n] + \lim \mathbb{E}[Y'_n] \quad \text{(properties of limits of real sequences)} \\
&= \mathbb{E}[X] + \mathbb{E}[X'] \quad \text{(MCT again, twice).}
\end{aligned}$$

Easy extension: $\mathbb{E}$ is a linear operator: $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$.

Note: this proving strategy is incredibly useful in practice, in part thanks to how simple the statement of the monotone convergence theorem is. Similar statements with Riemann integrals are not as simple. Just this proving method arguably justifies the effort of learning to be a user of measure theory.

3.8 Integrals of random variables taking negative values

For a general random variable X:



1. Write $X = X^+ + X^-$, where X^+ and X^- are the negative and positive parts respectively. For example, $X^- = \mathbf{1}[X < 0]X$.
2. Note: $-X^-$ is non-negative.
3. Compute $I^+ \coloneqq \mathbb{E}[X^+]$ and $I^- \coloneqq \mathbb{E}[-X^-]$.
4. If both $I^+ = I^- = \infty$, return an error ("the Lebesgue integral is not defined"),
5. else define $\mathbb{E}[X] \coloneqq I^+ - I^-$.

Terminology:

- If at least one of I^+ and I^- is finite, we say the Lebesgue integral of X is defined,
- when both I^+ and I^- are finite, we say X is *integrable*, denoted $X \in \mathbf{L}_1$.

Exercise: find a random variable $X \notin \mathbf{L}_1$ such that $\mathbb{E}X$ exists. Find a random variable Y where $\mathbb{E}Y$ is not defined.

3.9 Integrals with respect to a measure

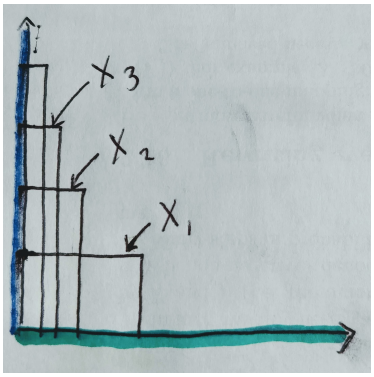
So far, we have assumed that the Lebesgue integral was computed with respect to probability measure $\mathbb{P} : \mathcal{F} \to [0, 1]$.

Exercise: go over the above argument again with a measure $\mu : \mathcal{F} \to [0, \infty)$ instead of a probability measure $\mathbb{P}$ and check that everything goes through.

3.10 More on exchanging limits and integrals

The monotone convergence theorem (Section 3.7) says that we can exchange integrals and limits *when the sequence of function is increasing* ($X_1 \leq X_2 \leq \dots$). Is this necessary?

Example showing that it is: consider $X_n = n\mathbf{1}_{(0,1/n]}$.



Exercise: compute $\lim X_n$ and $\mathbb{E}X_n$. Conclude that limits and integrals cannot be exchanged in this case.

However: there are non-monotone cases where you can exchange limits and integrals. For example, when they have **an integrable envelope**, defined as a random variable such that:

1. $|X_i| \leq Y$,
2. $\mathbb{E}|Y| < \infty$.

Theorem: (Dominated Convergence Theorem, DCT) if $X_1, X_2, \ldots$ have an integrable envelope, and $\lim X_i$ exists, then $\lim \mathbb{E}X_i = \mathbb{E} \lim X_i$.

3.11 Measure zero sets and almost sure statements

As a direct consequence of the axioms of probability, an empty event has probability zero: $\mathbb{P}(\emptyset) = 0$. The converse is not true though. For example under the uniform probability an event that contains only a single point still has probability zero $\mathbb{P}(\{0.2\}) = 0$. In fact, under the uniform probability, events containing countably many points have probability zero.[5]

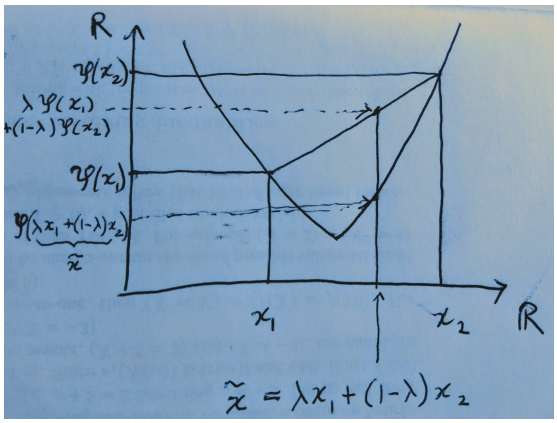
For this reason, it is often possible to relax a statement like "$|X_i| \leq Y$" in the previous statement to a statement like "$|X_i| \leq Y$ except for a set of probability zero." This second statement is formalized as $\mathbb{P}(|X_i| \leq Y \text{ for all } i) = 1$, and denoted "$|X_i| \leq Y$ a.s."

3.12 Convexity and integration

Review: convexity. A function $g : \mathbb{R} \to \mathbb{R}$ is convex if for all $x_1, x_2 \in \mathbb{R}$ and $\lambda \in [0, 1]$,

$$\lambda g(x_1) + (1 - \lambda)g(x_2) \geq g(\lambda x_1 + (1 - \lambda)x_2).$$

[5]Even more surprising, there are set containing uncountably many points that still have probability zero. Read about the Cantor set if you are curious.



Exercise: convince yourself that $\varphi(\mathbb{E}X) \neq \mathbb{E}[\varphi(X)]$ in general (with some exceptions to this, e.g. when φ is linear). This is unfortunate because $\varphi(\mathbb{E}X)$ is often easier to compute than $\mathbb{E}(\varphi(X))$.

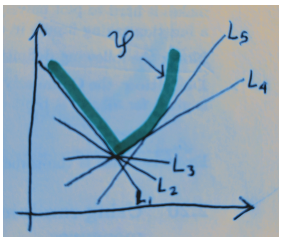
However: if φ is convex, we can at least get the following bound.

Jensen's inequality: if φ is convex, then $\varphi(\mathbb{E}X) \leq \mathbb{E}[\varphi(X)]$.

Proof: to prove Jensen's inequality, we will use the following result from convex analysis:

Lemma: for all convex function φ, there is a sequence of linear functions $L_n(x) = a_n x + b_n$ such that

$$\varphi(x) = \sup_n L_n(x).$$



Exercise: provide an example showing that the above lemma does not hold in the case for non-convex functions.

Back to Jensen's: using our lemma, we have $\varphi(X) \geq L_n(X)$ by construction. "Taking expectations on both sides" (i.e. using monotonicity of expectations):

$$\begin{aligned} \mathbb{E}\varphi(X) &\geq \mathbb{E}[L_n(X)] && (1)\\ &= L_n(\mathbb{E}X) \text{ since } L_n \text{ is linear.} && (2) \end{aligned}$$

Finally, taking sup over n on both sides:

$$\begin{aligned} \mathbb{E}[\varphi(X)] &\geq \sup L_n(\mathbb{E}X) && (3)\\ &= \varphi(\mathbb{E}X). && (4) \end{aligned}$$

Example: since $\varphi(x) = x^2$ is convex, Jensen's gives us the following inequality: $(\mathbb{E}X)^2 \leq \mathbb{E}[X^2]$.

Application: the variance is defined as $\mathbf{Var}[X] \coloneqq \mathbb{E}[(X - \mu)^2]$, where $\mu = \mathbb{E}X$. By computing the square and linearity, this last equation is equal to $\mathbb{E}[X^2] - (\mathbb{E}X)^2$. Therefore, Jensen's inequality gives us another proof that the variance is non-negative.

3.13 Markov's inequality and its friends

Motivation: let X be the water level near a dam of height of 7.5m. What is the probability of a flood? All you know is that the mean water level is 5m.

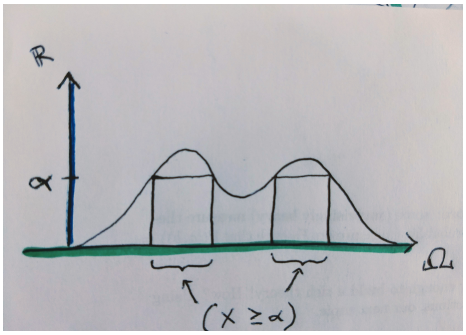
Proposition (Markov's inequality): if $X \geq 0$, then for all $\alpha \geq 0$,

$$\mathbb{P}(X \geq \alpha) \leq \frac{\mathbb{E}X}{\alpha}.$$

Exercise: solve the dam problem using Markov's inequality.

Proof: Convince yourself by looking at the graph of X that the following identities hold:

$$\begin{aligned} X &\geq \mathbf{1}[X \geq \alpha]X \quad \text{(follows from } X \geq 0\text{)} \\ &\geq \alpha \mathbf{1}[X \geq \alpha]. \end{aligned}$$



Taking expectations on both sides:

$$\begin{aligned} \mathbb{E}X &\geq \alpha \mathbb{E}[\mathbf{1}[X \geq \alpha]] \\ &= \alpha \mathbb{P}(X \geq \alpha). \end{aligned}$$

Note: non-negativity is necessary. Consider for example a random variable with the discrete uniform distribution on $\{-1, +1\}$, $\alpha = 1$.

Note: the bound will sometimes be greater than one. On the other hand, there are random variables X and α such that the bound it tight, i.e. where $\mathbb{P}(X \geq \alpha) = \frac{\mathbb{E}X}{\alpha}$. For example X such that $\mathbb{P}(X = 2) = 1/2$, $\mathbb{P}(X = 0) = 1/2$, $\alpha = 2$.

Corollary 1: for all random variable X, $p \geq 0$,

$$\mathbb{P}(|X| \geq \alpha) \leq \frac{\mathbb{E}|X|^p}{\alpha^p}.$$

Proof: since the power function is strictly increasing, $(|X| \geq \alpha) = (|X|^p \geq \alpha^p)$. Why? Manipulations of this kind are explained in gory details in the next few section, see specifically 3.14 for this step. Therefore $\mathbb{P}(|X| \geq \alpha) = \mathbb{P}(|X|^p \geq \alpha^p)$. Since $|X|^p \geq 0$, we can apply Markov's inequality on this last probability.

Corollary 2 (Chebyshev's inequality): for any random variable Y with $\mu \coloneqq \mathbb{E}Y$, $|\mu| < \infty$ (non-negativity not needed anymore!),

$$\mathbb{P}(|Y - \mu| \geq \alpha) \leq \frac{\mathbf{Var}Y}{\alpha^2}.$$

Proof: define $X \coloneqq |Y - \mu|$, and apply the preceding lemma with $p = 2$.

Corollary 3 (Chernoff bound): if X is any random variable, then

$$\mathbb{P}(X \geq \alpha) \leq \inf_t \frac{\mathbb{E}[e^{tX}]}{e^{t\alpha}}.$$

Proof: For any fixed t, the function $f(x) = e^{tx}$ is strictly increasing, so we can use the same reasoning as Chebyshev's. Since the inequality if true for all t, the left-hand side is bounded by the infimum over t.

Note: the quantity $\mathbb{E}[e^{tX}]$, viewed as a function of t, is called the *moment generating function*. We will meet this object again.

We will also revisit Chernoff's bound soon once we have introduced independence of random variables.

3.14 Rewriting events involving equalities

This section and the next two provide additional details as well as generalize one of the key step used in the proof of Markov's, Chebyshev's, and Chernoff's inequalities.

Some background: when dealing with equations involving non-random variables, a common heuristic is to "add on both sides," e.g. $x + 5 = 2$ becoming $x = -3$. Let us call these two logical statements s_1 and s_2. Since $s_1(X(\omega))$ is true if and only if $s_2(X(\omega))$ is true, it follows that the two events, $(X + 5 = 2)$ and $(X = -3)$, are equal. In particular, $\mathbb{P}(X + 5 = 2) = \mathbb{P}(X = -3)$.

More generally, if f is one-to-one, then $(X = Y) = (f(X) = f(Y))$. For example $(-X = -5) = (X = 5)$.

In fact, we only need f to be one-to-one on the set of possible values attained by X and Y (i.e. the union of their ranges). For example, if X is non-negative (denoted $X \geq 0$), then $(X = 2) = (X^2 = 4)$. Indeed, even though $f(x) = x^2$ is not one-to-one, it is when restricted to the positive real line.

Many steps in probability arguments follow that kind of logic-based reasoning.

3.15 Rewriting events involving inequalities

Be a bit more careful when dealing with inequalities, we need a more specialized form a one-to-one mapping: if f is *strictly increasing*, then $(X \leq Y) = (f(X) \leq f(Y))$. For example $(X \leq 0) = (\exp(X) \leq 1)$.

This is indeed necessary as $(-X \leq -5) = (X \geq 5) \neq (X \leq 5)$.

3.16 Bounding events

When we cannot rewrite the event using the methods described in Sections 3.14 and 3.15, we can settle on bounding the event instead. Suppose event A_1 is defined using statement $s_1(Y)$ (e.g. $A_1 = (Y = 2)$) and similarly, event A_2, using $s_2(Y)$ (e.g. $A_2 = (Y^2 = 4)$. Suppose $s_1 \implies s_2$. Then we have $A_1 \subseteq A_2$. In our example, $y = 2 \implies y^2 = 4$, but the converse is not true if the random variable Y can take negative values, as $y = -2$ also yields $y^2 = 4$. So all we get for general random variables is that $(Y = 2) \subset (Y^2 = 4)$.

Now we have seen in Section 2.24 that $A_1 \subseteq A_2$ implies $\mathbb{P}(A_1) \leq \mathbb{P}(A_2)$.

4 Independence

4.1 More than one random variables (random vectors)

Motivation: A man and a woman try to meet at a certain place between 1:00pm and 2:00pm. Suppose each person pick an arrival time between 1:00pm and 2:00pm uniformly at random (denote X and Y respectively), and waits for the other at most 10 minutes. What is the probability that they meet?

Practical question: How to compute a probability of the form $\mathbb{P}((X, Y) \in S)$?

Theoretical question: Given two random variables on the same space, $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$, what is (X, Y) exactly?

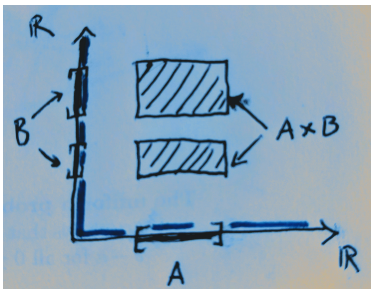
Definition: A random vector is a random variable that takes values in $\mathcal{X} = \mathbb{R}^2$:

$$(X, Y) : \Omega \to \mathbb{R}^2.$$

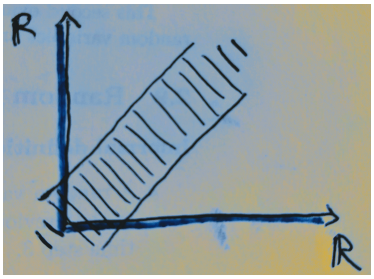
Recall: The definition of random variable requires a σ-algebra on $\mathcal{X}$. What should we pick? We know that we will at the very least need to compute the probability of X falling in a *rectangle*:

Definition: a rectangle is a set of the form:

$$\begin{aligned} R &= A \times B \text{ for some } A \in \mathcal{F}_{\mathrm{B}}, B \in \mathcal{F}_{\mathrm{B}} && (5) \\ &= \{(x, y) : x \in A, y \in B\}. && (6) \end{aligned}$$



Unfortunately: $\mathcal{S} = \{R : R \text{ is a rectangle}\}$ is not a σ-algebra (why?). Also, it does not contain the set that we would need to answer the "practical question" that kicked off this section.



Solution: generate a σ-algebra from $\mathcal{S}$. The result of this is called the product σ-algebra:

$$\mathcal{F}^{\otimes 2} \coloneqq \mathcal{F} \otimes \mathcal{F} \coloneqq \sigma(\mathcal{S}).$$

Exercise: show that the set S in the "practical question" is in $\mathcal{F} \otimes \mathcal{F}$.

4.2 Distribution, CDF and density of a random vector

These definitions follow directly from the univariate case:

Definition: the joint distribution of a random vector (X, Y), denoted $\mathbb{P}_{X,Y} : \mathcal{F} \otimes \mathcal{F} \to [0, 1]$, is defined by:

$$\mathbb{P}_{X,Y}(S) \coloneqq \mathbb{P}((X, Y) \in S).$$

Definition: the joint CDF of a random vector (X, Y), denoted $F_{X,Y} : \mathbb{R}^2 \to [0, 1]$, is defined by:

$$F_{X,Y}(x, y) \coloneqq \mathbb{P}_{X,Y}((-\infty, x] \times (-\infty, y]).$$

Note: This last definition seems slightly arbitrary. Why pick this particular class of sets S ("quarter-planes")? As we will see in the next section, this is because this class *characterizes* the joint distribution. In other words, given a joint CDF, you can in principle obtain the probability of X falling in any set S.

Definition: a function $f : \mathbb{R}^2 \to [0, \infty)$ is called a joint density of the random vector (X, Y) if:

$$F_{(X,Y)}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(x, y) \, \mathrm{d}x \, \mathrm{d}y.$$

Note: these definitions can be generalized to more than two dimensions.

4.3 Determining classes

In this section, we explain the tool used to prove the characterization statement of the previous section.

Tool: π-λ theorem.

1. Let π denote a collection of sets satisfying the following condition (called a π-system condition):
   (a) $A, B \in \pi \implies A \cap B \in \pi$.
2. Let λ denote a collection of sets satisfying the following conditions (called a λ-system condition):
   (a) $\Omega \in \lambda$,
   (b) $A, B \in \lambda, A \subseteq B \implies B \backslash A \in \lambda$,
   (c) $A_i \uparrow A, A_i \in \lambda \implies A \in \lambda$.
3. Then, the following holds: $\pi \subset \lambda \implies \sigma(\pi) \subset \lambda$.

Proposition: Supposet $\mathbb{P}_1$ and $\mathbb{P}_2$ are probability measures on $\mathcal{F} = \sigma(\mathcal{S})$, where $\mathcal{S}$ is closed under finite intersection. Then the following are equivalent:

1. $\mathbb{P}_1(A) = \mathbb{P}_2(A)$ for all $A \in \mathcal{F}$,
2. $\mathbb{P}_1(A) = \mathbb{P}_2(A)$ for all $A \in \mathcal{S}$.

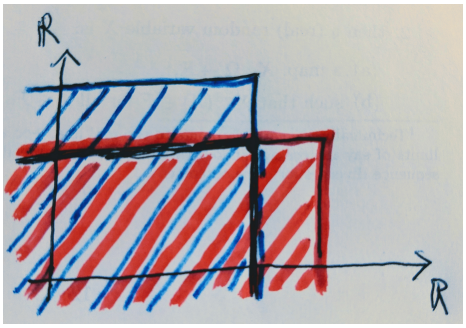
Exercise: prove this using the π-λ theorem. Hint: let $\lambda = \{A \in \mathcal{F} : \mathbb{P}_1(A) = \mathbb{P}_2(A)\}$.

Corollary: F_X determines $\mathbb{P}_X$.

Proof: $(-\infty, x] \cap (-\infty, y] = (-\infty, \min\{x, y\}]$. Hence $\mathcal{S} = \{(-\infty, x]\}$ is a π system. Conclude with the π-λ theorem.

Corollary: $F_{X,Y}$ determines $\mathbb{P}_{X,Y}$.

Proof: Use the same proof strategy, this time based on the following picture:



Corollary: if (X, Y) has joint density f, then

$$\mathbb{P}((X, Y) \in S) = \int_S f(x, y)\,\mathrm{d}x\,\mathrm{d}y.$$

4.4 Independence of random variables

Definition: the random variables $X_1, X_2, \ldots, X_n : \Omega \to \mathbb{R}$ are independent if, for all $A_i \in \mathcal{F}_{\mathrm{B}}$,

$$\mathbb{P}(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i).$$

Equivalent notation:

$$\mathbb{P}_{X_1, X_2, \ldots, X_n}(A_1 \times A_2 \times \cdots \times A_n) = \prod_{i=1}^{n} \mathbb{P}_{X_1}(A_i).$$

Exercise: using $\pi - \lambda$ show that the above statement can be checked by showing that the CDF factorizes as

$$F_{X_1, X_2, \ldots, X_n} = \prod_{i=1}^{n} F_{X_i}.$$

Note: similarly to the above exercise, if the random variables have a joint density $f_{X_1, X_2, \ldots, X_n}$, independence can be checked by factorizing the density into *marginal densities* f_{X_i} of each individual random variable:

$$f_{X_1, X_2, \ldots, X_n} = \prod_{i=1}^{n} f_{X_i}.$$

Definition: we say the random variables $X_1, X_2, \ldots, X_n$ are pairwise independent if each pair is independent.

Exercise: find 3 random variables such that these random variables are pairwise independent but not independent. Hint: this can be done using indicator random variables.

4.5 Chernoff's bound, continued

We encountered the moment generating function $f(t) = \mathbb{E}[e^{tX}]$ when talking about Chernoff's bound. When X is a sum of independent random variables, $X = X_1 + X_2 + \cdots + X_n$, we get this nice formula:

$$f(t) = \prod_{i=1}^{n} f_i(t),$$

where $f_i(t) = \mathbb{E}[e^{tX_i}]$. Informally, adding random variables multiplies their moment generating functions.

Since the moment generating function for individual variables are often easy to compute, this gives a powerful way to evaluate the right hand side of Chernoff's bound

Example: Chernoff-Hoeffding theorem if the X_i are independent and each have a Bernoulli(p) distribution (terminology: "are iid Bernoulli"), and $\epsilon > 0$:

$$\mathbb{P}\left(\frac{1}{n}\sum X_i \geq p + \varepsilon\right) \leq \left(\left(\frac{p}{p+\varepsilon}\right)^{p+\varepsilon}\left(\frac{1-p}{1-p-\varepsilon}\right)^{1-p-\varepsilon}\right)^n$$
$$\mathbb{P}\left(\frac{1}{n}\sum X_i \leq p - \varepsilon\right) \leq \left(\left(\frac{p}{p-\varepsilon}\right)^{p-\varepsilon}\left(\frac{1-p}{1-p+\varepsilon}\right)^{1-p+\varepsilon}\right)^n$$

Proof: compute the moment generating function of a Bernoulli. Then use calculus to minimize over t. Then plug-in the minimum. Use the variable $Y_i = 1 - X_i$ to establish the other side of the bound.

Other versions: see the wikipedia page for many more bounds derived from this.

4.6 Exercise set 3

1. Exercise in Section 3.4 about linearity in the case of simple functions.
2. Solve the two exercises in Section 3.12.
3. Motivating exercise in Section 3.13.
4. Exercise in Section 4.3.

4.7 Solutions for exercise set 3

1. Let $Y = \sum_{i=1}^{N} a_i \mathbf{1}_{A_i}$ and $Y' = \sum_{j=1}^{M} b_j \mathbf{1}_{B_j}$ for $\{A_i\}_{i=1}^N, \{B_j\}_{j=1}^M$ are measurable disjoint collections. Notice that since $\{A_i\}_{i=1}^N$ and $\{B_j\}_{j=1}^M$ are disjoint, we have

$$\mathbf{1}_{A_i} = \sum_{j=1}^{M} 1_{A_i \cap B_j}, \quad \mathbf{1}_{B_j} = \sum_{i=1}^{N} 1_{A_i \cap B_j}.$$

Therefore, $Y + Y'$ can be written as

$$\begin{aligned} Y + Y' &= \sum_{i=1}^{N} a_i \mathbf{1}_{A_i} + \sum_{j=1}^{M} b_j \mathbf{1}_{B_j} \\ &= \sum_{i=1}^{N} a_i \sum_{j=1}^{M} 1_{A_i \cap B_j} + \sum_{j=1}^{M} b_j \sum_{i=1}^{N} 1_{A_i \cap B_j} \\ &= \sum_{i=1}^{N} \sum_{j=1}^{M} (a_i + b_j) 1_{A_i \cap B_j}. \end{aligned}$$

Since $A_i \cap B_j$ are disjoint, we have $Y + Y'$ is a simple function and by the definition of the expectation,

$$\begin{aligned} \mathbb{E}[Y + Y'] &= \sum_{i=1}^{N} \sum_{j=1}^{M} (a_i + b_j) \mathbb{P}(A_i \cap B_j) \\ &= \sum_{i=1}^{N} a_i \sum_{j=1}^{M} \mathbb{P}(A_i \cap B_j) + \sum_{j=1}^{M} b_j \sum_{i=1}^{N} \mathbb{P}(A_i \cap B_j) \\ &= \sum_{i=1}^{N} a_i \mathbb{P}(A_i) + \sum_{j=1}^{M} b_j \mathbb{P}(B_j) \\ &= \mathbb{E}[Y] + \mathbb{E}[Y']. \end{aligned}$$

2. (a) Let $\Omega = [0, 1]$ with $\mathbb{P}$ be the Lebesgue measure and $X : \Omega \to \mathbb{R}$ be the identity. If $\varphi : \mathbb{R} \to \mathbb{R}$ be $\varphi(x) = x^2$. Then,

$$\mathbb{E}(X) = \int_{[0,1]} x \mathbb{P}(dx) = \int_0^1 x dx = \frac{1}{2},$$

and similarly,

$$\mathbb{E}(X^2) = \int_{[0,1]} x^2 \mathbb{P}(dx) = \int_0^1 x^2 dx = \frac{1}{3}.$$

Therefore $\varphi(\mathbb{E}[X]) = 1/4$ which does not equal $\mathbb{E}[\varphi(X)] = 1/3$.

(b) Let $\varphi = \mathbf{1}_{\{0\}}$. Suppose that there is a sequence of linear functions $L_n(x) = a_n x + b_n$ such that $\sup_n L(x) = \varphi(x)$ for all x. Suppose that $a_{n_0} \neq 0$ for some n_0. Then L_{n_0} is unbounded from above and there is some $x_0 \in \mathbb{R}$ such that $L_{n_0}(x_0) = 2$. Thus $\sup_n L_n(x_0) \geq 2 > \varphi(x_0)$ which is a contradiction. Thus we have all $a_n = 0$.

For $x \neq 0$, we have $\varphi(x) = 0 = \sup_n L_n(x) = \sup_n b_n$, which implies $b_n \leq 0$ for all n. However when $x = 0$, and $\sup_n L_n(0) = \sup_n b_n \leq 0 \neq 1\varphi(x)$.

Thus we have shown there does not exist a sequence of linear functions L_n such that $\sup_n L(x) = \varphi(x)$ for all x.

3. There is a flood when the water level X is higher than the dam, which is at 7.5m. By Markov's inequality,

$$\mathbb{P}(\text{Flood}) = \mathbb{P}(X \geq 7.5) \leq \frac{\mathbb{E}X}{7.5} = \frac{5}{7.5} = \frac{2}{3}.$$

Therefore a flood will occur with maximum probability 2/3.

4. "(1) $\to$ (2)." This follows trivially since $\mathcal{S} \subset \mathcal{F}$.

"(2) $\to$ (1)." As suggested by the hint, we define

$$\lambda = \{A \in \mathcal{F} : \mathbb{P}_1(A) = \mathbb{P}_2(A)\}.$$

We will now show that λ is a λ-system:

1. Note that $\mathbb{P}_1(\Omega) = 1 = \mathbb{P}_2(\Omega)$, thus $\Omega \in \lambda$.
2. Suppose $A, B \in \lambda$ and $A \subset B$. We then have,

$$\begin{aligned} \mathbb{P}_1(B \backslash A) &= \mathbb{P}_1(B) - \mathbb{P}_1(A) \\ &= \mathbb{P}_2(B) - \mathbb{P}_2(A) \text{ (since } A, B \in \lambda) \\ &= \mathbb{P}_2(B \backslash A). \end{aligned}$$

3. Suppose $A_i \uparrow A$ for all $A_i \in \lambda$. Then

$$\begin{aligned} \mathbb{P}_1(A) &= \lim_{i \to \infty} \mathbb{P}_1(A_i) \text{ (continuity of measures)} \\ &= \lim_{i \to \infty} \mathbb{P}_2(A_i) \text{ (since } A_i \in \lambda) \\ &= \mathbb{P}_2(A) \text{ (continuity of measures).} \end{aligned}$$

Therefore we have shown that λ is a λ-system. Since $\mathcal{S}$ is closed under finite intersection, we have by definition it is a π-system. Thus by the $\lambda - \pi$ theorem $\mathcal{F} = \sigma(\mathcal{S}) \subset \lambda$, which completes the proof.

5 Computing expectations in practice

5.1 Computing integrals using calculus

When $\Omega = [a, b] \subset \mathbb{R}$, μ is uniform, and f is bounded, then the Lebesgues integral $\int f \, \mathrm{d}\mu$ behaves in the same way as the standard Riemann integral (as long as the latter exists, which is true iff the set of discontinuities of f have measure zero). Hence you can in principle use the **fundamental theorem**

of calculus to compute these integrals: if there is a function F such that $F'(x) = f(x)$ for all $x \in \Omega$, then

$$\int f \, \mathrm{d}\mu = \int f(x) \, \mathrm{d}x = F(b) - F(a).$$

It also follows that F is the CDF of X.

The limitation of this approach is that even when f is a simple close-form expression (expressed in terms of $+, -, *, /, \exp, \log$), the CDF F might not have such a close-form expression (example: $f(x) = \exp(-x^2)$).

5.2 Computing expectations using the distribution of the random variable

Motivating example: consider a space Ω containing the following four objects: a circle, a triangle, a square, and a pentagon, and a probability $\mathbb{P}$ that give then equal probabilities $(1/4)$. Let X denote a random variable that takes a shape as input and output the number of faces. Suppose we want to compute $\mathbb{E}[g(X)]$, where $g(x) = \mathbf{1}[x \text{ is an even integer}]$. We will see two methods for doing this:

From the definition: let $Y = g(X)$. We see that Y can take two possible values, zero and one, therefore it is simple. Applying the definition of expectation of simple function:

$$\begin{aligned}
\mathbb{E}[Y] &= \int Y \, \mathrm{d}\mathbb{P} \\
&= 1 \times \mathbb{P}(Y = 1) + 0 \times \mathbb{P}(Y = 0) \\
&= \mathbb{P}(Y^{-1}(\{1\}) \\
&= \mathbb{P}(X^{-1}(g^{-1}(\{1\}))) = 1/2.
\end{aligned}$$

Another way: first, derive the distribution of X:

$$\mathbb{P}_X(A) = \frac{1}{4} \left(\mathbf{1}[0 \in A] + \mathbf{1}[3 \in A] + \mathbf{1}[4 \in A] + \mathbf{1}[5 \in A] \right).$$

then, compute an integral of g *with respect to the distribution of* X:

$$\int g \, \mathrm{d}\mathbb{P}_X.$$

Here, g is an indicator, so it is a simple function, so we can use our definition of integral of simple functions:

$$\begin{aligned}
\int g \, \mathrm{d}\mathbb{P}_X &= 1 \times \mathbb{P}_X(\{x : g(x) = 1\}) + 0 \times \mathbb{P}_X(\{x : g(x) = 0\}) \\
&= \mathbb{P}_X(\{\dots, -6, -4, -2, 0, 2, 4, 6, \dots\}) \\
&= 1/2.
\end{aligned}$$

Proposition ("change of variable"): these two methods are equivalent. More precisely, if $Y = g(X)$ is a random variable and either $g \geq 0$ or $Y \in \mathbf{L}_1$, then:

$$\int Y \, \mathrm{d}\mathbb{P} = \int g \, \mathrm{d}\mathbb{P}_X.$$

Note: by "either $g \geq 0$ or $Y \in \mathbf{L}_1$" I mean that the above can actually be slit into two propositions, one assuming the first condition, and a second proposition assuming the second condition.

Note: an important special case is $g(x) = x$.

Note: the nice thing with the second way is that you do not have to know Ω and $\mathbb{P}$, which are often not explicitly given to you. Often all I tell you is $X \sim F$, which characterizes $\mathbb{P}_X$, and I ask $\mathbb{E}[g(X)]$. Using our proposition we can solve this using an integral over the real line with a measure on the reals provided by $\mathbb{P}_X$.

Proof: assume first that $g = \mathbf{1}_A$. From the same computation as in the example above, we have on one hand:

$$\int Y \, \mathrm{d}\mathbb{P} = \mathbb{P}(Y = 1),$$

and on the other hand:

$$\begin{aligned} \int g \, \mathrm{d}\mathbb{P}_X &= \mathbb{P}_X(\{x : g(x) = 1\}) \\ &= \mathbb{P}(X \in \{x : g(x) = 1\}) \\ &= \mathbb{P}(g(X) = 1) \\ &= \mathbb{P}(Y = 1). \end{aligned}$$

Next, assume g is simple, i.e. $g = a_1 \mathbf{1}_{A_1} + \cdots + a_n \mathbf{1}_{A_n}$. We have:

$$\begin{aligned} \int g(X) \, \mathrm{d}\mathbb{P} &= \int \sum_i a_i \mathbf{1}_{A_i}(X) \, \mathrm{d}\mathbb{P} \\ &= \sum_i a_i \int \mathbf{1}_{A_i}(X) \, \mathrm{d}\mathbb{P} \quad \text{(using linearity)} \\ &= \sum_i a_i \int \mathbf{1}_{A_i} \, \mathrm{d}\mathbb{P}_X \quad \text{(using first part of proof)} \\ &= \int \left(\sum_i a_i \mathbf{1}_{A_i} \right) \mathrm{d}\mathbb{P}_X \quad \text{(using linearity)} \\ &= \int g \, \mathrm{d}\mathbb{P}_X. \end{aligned}$$

Third, assume $g \geq 0$ (or $g \in \mathbf{L}_1$). Use the approximation theorem to get $g_i \uparrow g$,

so that:

$$\begin{aligned}
\int g(X)\,\mathrm{d}\mathbb{P} &= \int \lim g_i(X)\,\mathrm{d}\mathbb{P} \\
&= \lim \int g_i(X)\,\mathrm{d}\mathbb{P} \text{ (using MCT (or DCT))} \\
&= \lim \int g_i\,\mathrm{d}\mathbb{P}_X \text{ (using second part of proof)} \\
&= \int \lim g_i\,\mathrm{d}\mathbb{P}_X \text{ (using MCT)} \\
&= \int g\,\mathrm{d}\mathbb{P}_X.
\end{aligned}$$

.

5.3 How probability spaces and random variables are constructed in practice

So far we have often insisted in explicitly constructed Ω, $\mathcal{F}$, $\mathbb{P}$, and designed random variables X by providing, for each ω, the value $X(\omega)$. This level of detail is useful for certain proofs (e.g. Markov's inequality), but for many day-to-day calculations, it is not necessary to go into that much detail.

Idea: instead of defining Ω, $\mathcal{F}$, $\mathbb{P}$ and X, just specify the distribution of X. I.e. say something like "let Ω, $\mathcal{F}$, $\mathbb{P}$ and X be such that $X \sim \mathrm{Bern}(p)$, i.e. such that X has a Bernoulli p distribution."

How do we know that there are such Ω, $\mathcal{F}$, $\mathbb{P}$ and X? Thanks to the inverse CDF construction, introduced in Section 2.26.

This does not uniquely define Ω, $\mathcal{F}$, $\mathbb{P}$ and X: recall from Section 2.22 that there are several ways to build Ω, $\mathcal{F}$, $\mathbb{P}$ and X that yield the same distribution on X.

However: in light of the preceding Section (5.2), we only need the distribution in order to compute arbitrary expectations. So even though Ω, $\mathcal{F}$, $\mathbb{P}$ and X are not technically fully specified, they are constrained enough for our purpose.

For this reason, often we do not even mention Ω, $\mathcal{F}$, $\mathbb{P}$, and just assume there is a "global" probability space on which the random variables are defined based on their distributions.

5.4 Computing expectations using densities

Generalization of density: we say X has a density f with respect to μ (typically, μ is the uniform measure on $\mathbb{R}$, but other choices are possible), if:

$$\mathbb{P}_X(A) = \int \mathbf{1}_A f\,\mathrm{d}\mu \quad \text{for all } A \in \mathcal{F}_{\mathrm{B}}.$$

Note: this generalizes our previous definition of density given in Section 2.23:

$$\begin{aligned} F_X(b) - F_X(a) &= \mathbb{P}_X([a,b]) \\ &= \int_a^b f(x)\,\mathrm{d}x \ \text{ (by Section 5.1)} \end{aligned}$$

Therefore by letting $a \to -\infty$ we recover the formula from Section 2.23.

Critical question in first assignment: show that if $g \geq 0$ or $g \in \mathbf{L}_1$,

$$\mathbb{E}[g(X)] = \int f(x)g(x)\mu(\,\mathrm{d}x).$$

Why this is useful: because we can now compute expectations using calculus, even when $\Omega \neq \mathbb{R}$, and by knowing only $\mathbb{P}_X$.

5.5 Computing the expectation of a function of independent random variables

Motivation: back to the motivating problem of Section 4.1. How to formalize the compution of the expectation, $\mathbb{E}[\mathbf{1}[|X-Y| \leq 1/6]]$. Let us define $g(x,y) \coloneqq \mathbf{1}[|x-y| \leq 1/6]$. By our change of variable formula (Section 5.2),

$$\mathbb{E}[g(X,Y)] = \int g\,\mathrm{d}\mathbb{P}_{X,Y}.$$

Now our independence assumption on the arrival times of the woman and man means that $\mathbb{P}_{X_1,X_2}(A_1 \times A_2) = \mathbb{P}_{X_1}(A_1)\mathbb{P}_{X_2}(A_2)$. To solve this last integral, we use this independence statement and the following theorems:

Tonelli: if $g \geq 0$, and $\mu \times \mu$ is such that $\mu \times \mu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$ (think of μ_1 and μ_2 as the marginal distributions of two *independent* random variables, i.e. $\mu \times \mu$ as $\mathbb{P}_{X,Y}$, μ_1 as $\mathbb{P}_X$ and μ_2 as $\mathbb{P}_Y$), then:

$$\begin{aligned} \int g\,\mathrm{d}(\mu_1 \times \mu_2) &= \int \left(\int g\,\mathrm{d}\mu_1 \right) \mathrm{d}\mu_2 \\ &= \int \left(\int g\,\mathrm{d}\mu_2 \right) \mathrm{d}\mu_1, \end{aligned}$$

in other words, we can approach the problem as iterated univariate integrals, and do so in any order we wish.

Fubini: same statement as Tonelli, but instead of requiring $g \geq 0$, we ask that $\int |g|\,\mathrm{d}(\mu_1 \times \mu_2) < \infty$. How to check this $\mathbf{L}_1$ condition? By first applying Tonelli on $|g| \geq 0$!
Being able to write joint integrals as iterated ones is nice because the inner integral is just over the real line, so we can use our density f of X_1 and calculus at this point.

Corollary: X and Y are independent if and only if for all measurable $g_i \geq 0$,

$$\mathbb{E}[g_1(X)g_2(Y)] = \mathbb{E}[g_1(X)]\mathbb{E}[g_2(Y)].$$

Solving the meeting problem: by Fubini, we have

$$\int g\mathbb{P}_{X,Y} = \int \left(\int g \, \mathrm{d}\mathbb{P}_X \right) \mathrm{d}\mathbb{P}_Y.$$

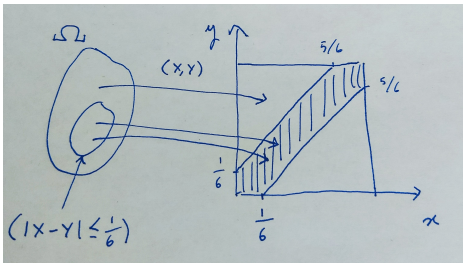
Now the inner integral can be rewritten as an integral over $[0,1]$ using the density question in the first assignment:

$$\int \left(\int g \, \mathrm{d}\mathbb{P}_X \right) \mathrm{d}\mathbb{P}_Y = \int \left(\int f(x)g(x,y) \, \mathrm{d}x \right) \mathrm{d}\mathbb{P}_Y,$$

where $f(x)$ is the uniform density over $[0,1]$, $f(x) = \mathbf{1}_{[0,1]}(x)$. Applying the same argument to the outer integral, we get:

$$\int \left(\int f(x)g(x,y) \, \mathrm{d}x \right) \mathrm{d}\mathbb{P}_Y = \int f(y) \int f(x)g(x,y) \, \mathrm{d}x \, \mathrm{d}y,$$

which is the area of the dashed region in the following figure:

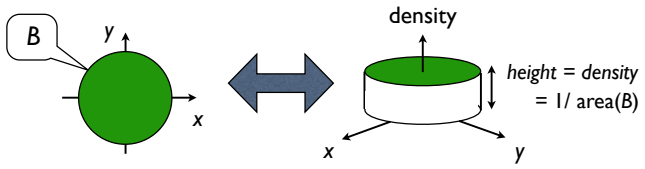


Important: the notion of *independent random variables* should not be confused with *uncorrelated random variables*: X and Y are uncorrelated if and only if

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

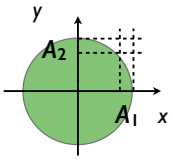
Note: X and Y independent implies they are uncorrelated,[6] but the converse is not true!

Exercise: consider (X,Y) with a uniform density on the unit circle.



[6] Assuming $XY \in \mathbf{L}_2$.

1. Find $\mathbb{E}[XY], \mathbb{E}[X], \mathbb{E}[Y]$. Hint: use symmetry.
2. Find g_i's such that $\mathbb{E}[g_1(X)g_2(Y)] = 0$ but $\mathbb{E}[g_1(X)]\mathbb{E}[g_2(Y)] > 0$. Hint: use the indicators shown in the figure below.



5.6 Declaring independent random variables

Let us continue the discussion of Section 5.3, on implicit specification of probability spaces. Suppose now we want to declare more than one random variables. Often we do so by declaring that (1) they are independent, and (2) the distribution of each random variable. For example: "let $X_1, X_2, \ldots$ be independent Bern(0.5) random variables." A shorthand: "$X_1, X_2, \ldots$ are i.i.d. Bern(0.5)" (identically and independently distributed). Or just:

$$X_i \overset{\text{iid}}{\sim} \text{Bern}(p).$$

There is always some Ω, $\mathbb{P}$, $\mathcal{F}$, X_1, X_2, ... satisfying the constraints (1) and (2): you can take this as a fact (search Kolmogorov's Extension Theorem if you are curious).

This does not uniquely specify Ω, $\mathbb{P}$, $\mathcal{F}$, X_1, X_2, ...: for the same reason as in Section 5.3.

But: in the light of Section 5.5, all we need to know to compute expectation is (1) to know that the random variables are independent and (2) the distribution of each random variable X_i (called the "marginal distributions").

5.7 Computing expectation using the cumulative distribution function

Idea: consider the following identity, which holds true for all $x \geq 0$,

$$x = \int_0^\infty \mathbf{1}[x > t] \, \mathrm{d}t.$$

Now for $X \geq 0$, let us take expectations on both sides and use Fubini:

$$\mathbb{E}X = \int_\Omega X \, \mathrm{d}\mathbb{P} = \int_\Omega \int_0^\infty \mathbf{1}[X > t] \, \mathrm{d}t \, \mathrm{d}\mathbb{P} = \int_0^\infty \int_\Omega \mathbf{1}[X > t] \, \mathrm{d}\mathbb{P} \, \mathrm{d}t,$$

which yields:

$$\mathbb{E}X = \int_0^\infty \mathbb{P}[X > t] \, \mathrm{d}t = \int_0^\infty (1 - F_X(t)) \, \mathrm{d}t.$$

5.8 Transformations of random variables and random vectors

One dimension. In light of exercise 6 in Section 2.28, it should not be a surprise how to solve the following problem: "given a random variable X with known density f_X (with respect to the Lebesgue measure), and strictly increasing, differentiable function g, find the density of $Y = g(X)$ (again, with respect to the Lebesgue measure)."

Solution: from exercise 6 in Section 2.28, we already know the CDF F_Y of Y, namely $F_Y(y) = F_X(g^{-1}(y))$. To get the density f_Y of Y, based on Section 5.1, we simply take the derivative with respect to y and use chain rule, obtaining:

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_X(g^{-1}(y)) = f_X(g^{-1}(y)) \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y).$$

The multivariate version is a natural extension of the above formula: let $X : \Omega \to \mathbb{R}^n$ denote a random vector with density $f_X : \mathbb{R}^n \to [0, \infty)$, let $g : \mathbb{R}^n \to \mathbb{R}^n$ be invertible with both g and g^{-1} continuously differentiable (a so called C_1 *diffeomorphism*), then the density of $f_Y : \mathbb{R}^n \to [0, \infty)$ is given by

$$f_Y(y) = f_X(g^{-1}(y)) \left| \det \left(\frac{\partial}{\partial y_c} g_r^{-1}(y) \right) \right|$$

where the derivative is generalized into the determinant of a Jacobian matrix indexed by rows $r \in \{1, 2, \ldots, n\}$ and columns $c \in \{1, 2, \ldots, n\}$.

Note: the above theorem can be modified to change $\mathbb{R}^n$ to open subsets of $\mathbb{R}^n$ in the obvious way, which is useful when X takes values in a subset of $\mathbb{R}^n$ such as the *simplex* (a list of n positive numbers that sum to one) in the case of a Dirichlet random variable.

For computations: it is often easier to compute the derivative of g rather than g^{-1}, and thankfully using properties of Jacobian matrices,

$$\left| \det \left(\frac{\partial}{\partial y_c} g_r^{-1}(y) \right) \right| = \frac{1}{\left| \det \left(\frac{\partial}{\partial y_c} g_r(y) \right) \right|}.$$

Example: if $X_1 \sim \mathrm{Exp}(1)$ and $X_2 \sim \mathrm{Exp}(1)$ are independent, show that $Y = X_1/(X_1 + X_2)$ is uniform between zero and one. Hint: the function $g(x_1, x_2) = x_1/(x_1 + x_2)$ is not a diffeomorphism (it looses information) so you cannot apply the above theorem directly! Use instead $g(x_1, x_2) = (x_1/(x_1 + x_2), x_1 + x_2)$. This augmentation trick is often useful. In computational statistics it forms the basis of "reversible jump Markov chain Monte Carlo," a method used in Bayesian statistics to select among models of different dimensionalities which fits better the data.

Do you really need to do this? Not always! Sometimes you may be only interested in the expectation of the transformed random variable $\mathbb{E}[Y]$. In this

case, we know from the first assignment that this can be done directly without computing g^{-1} or derivatives via:

$$\mathbb{E}[Y] = \int g(x) f_X(x) \, \mathrm{d}x.$$

This is often much quicker than finding f_Y and then computing

$$\mathbb{E}[Y] = \int y f_Y(y) \, \mathrm{d}y.$$

However in other cases, such as when a random variable is not well summarized by its first few moments (e.g. mean and variance), you will have to compute f_Y.

6 Asymptotics

Often we are faced with a question about a large number of random variables, say $X_1, X_2, \ldots, X_{1,000,000,000}$. In the context of big data, or big models, or large Monte Carlo simulations, the number of random variables of interest can be quite large. Providing an exact answer to questions involving so many random variables is often computationally prohibitive or impossible.

Key observation: paradoxically, under certain angles, the behaviour of large sets of random variables becomes increasingly simple. This allows us to make approximations. Asymptotics is the field concerned with making sure that these approximations can become arbitrarily accurate as the number of random variables increases. Moreover, asymptotics can sometimes give some hint on how to compare different approximations, i.e. how fast approximations converge (*rates*).

6.1 Infinitely often and eventually

Motivation: Let $X_1, X_2, X_3, \ldots$ be iid non-atomic (continuous) random variables representing the best performance achieved in a sport (e.g. 100m freestyle) at consecutive olympic games. Let R_n denote the indicator variable that a record is broken at olympic n (i.e. $R_n = 1$ if $X_n > X_j$ for all $j = 1, 2, \ldots, n-1$, and $R_n = 0$ otherwise). Let $A_n = (R_n = 1)$. What is the probability that records are broken infinitely often? One! Let $B_n = (R_n R_{n+1} = 1)$. What is the probability that records are broken in two consecutive years infinitely often? Zero! Why? And how to define "infinitely often" formally?

Definition: Let A_n denote an infinite collection of events, $n \in \{1, 2, 3, \ldots\}$ (not necessarily nested). We create two new events from these:

$$\begin{aligned} (A_n \text{ ev.}) &\coloneqq \{\omega \in \Omega : \exists N \in \{1, 2, \ldots\}, \forall n \geq N, \omega \in A_n\} \\ (A_n \text{ i.o.}) &\coloneqq \{\omega \in \Omega : \forall N \in \{1, 2, \ldots\}, \exists n \geq N, \omega \in A_n\}. \end{aligned}$$

Examples: consider the "drunk train" example from the first lecture.

- Let $X_i = 2Y_i - 1$ denote the direction the train takes at each step where $Y_i \sim \text{Bern}(1/2)$ are iid (independent and identically distributed)
- Let $S_n = X_1 + X_2 + \cdots + X_n$ denotes the current position of the train.
- Let A_n denote the event that the train is at "home" at step n, i.e. $A_n = (S_n = 0)$.
- The event that the train never gets lost: call this the event A, i.e. that the train returns to zero infinitely often. $A \coloneqq (A_n \text{ i.o.})$.
- The event that train eventually gets lost: $B \coloneqq (S_n \neq 0 \text{ ev.})$.

Proposition: $(A_n \text{ i.o.}) = (A_n^{\complement} \text{ ev.})^{\complement}$.

Proof: by definition:

$$\begin{aligned}(A_n \text{ i.o.}) &= \bigcap_{N=1}^{\infty} \bigcup_{n \geq N} A_n \\ (A_n \text{ ev.}) &= \bigcup_{N=1}^{\infty} \bigcap_{n \geq N} A_n.\end{aligned}$$

The proposition follows from De Morgan's law.

6.2 Borel-Cantelli (BC) lemma 1

Proposition: If

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty,$$

then $\mathbb{P}(A_n \text{ i.o.}) = 0$.

Notes:

- A_n do not need to be independent/disjoint/nested!
- But independence will be needed for a partial converse (BC 2)

Example: A drunk bird eventually gets lost (with probability one).

- Let $X_i^{(3)} \coloneqq (X_i^{(1)}, X_i'^{(1)}, X_i''^{(1)})$, where the three coordiates are indendent copies of the random walk of Section 6.1. I.e. diagonal moves are permitted.
- Similarly, $S_n^{(3)} = X_1^{(3)} + X_2^{(3)} + \cdots + X_n^{(3)}$.
- The claim can be rewritten as $\mathbb{P}(S_n^{(3)} = (0,0,0) \text{ i.o.}) = 0$.

1. First, by the binomial formula,

$$\mathbb{P}(S_{2n}^{(1)} = 0) = \binom{2n}{n} 2^{-2n}.$$

   Why? Well each individual path of length $2n$ has probability 2^{-2n}. How many such paths are at zero at step $2n$? Those where you go up n times and down n times. Counting the number of paths is like counting the subset of $\{1, 2, \ldots, 2n\}$ where we go up. There are $\binom{2n}{n}$ subsets of $\{1, 2, \ldots, 2n\}$ of size n.

2. Next, apply Stirling's formula ($n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$) to get

$$\mathbb{P}(S_{2n}^{(1)} = 0) \sim \frac{c}{\sqrt{n}},$$

   where c is a constant, and $a_n \sim b_n$ means that $a_n/b_n \to 1$.

3. This means that for the whole process,

$$\mathbb{P}(S_{2n}^{(3)} = 0) \sim \frac{c}{n^{3/2}},$$

   which is summable, i.e.

$$\sum_{n=1}^{\infty} \frac{c}{n^{3/2}} < \infty,$$

   because $3/2 > 1$.

4. Hence, by BC 1, $\mathbb{P}(S_n^{(3)} = (0,0,0) \text{ i.o.}) = 0$.

Proof of BC: let N denote the number of A_n's that occur:

$$N = \sum_{n=1}^{\infty} \mathbf{1}_{A_n}.$$

We have:

1. $(N = \infty) = (A_n \text{ i.o.})$,

2. By MCT:

$$\begin{aligned} \mathbb{E}[N] &= \mathbb{E}\left[\sum_{n=1}^{\infty} \mathbf{1}_{A_n}\right] \\ &= \sum_{n=1}^{\infty} \mathbb{P}(A_n). \end{aligned}$$

3. Hence by the assumption, $\mathbb{E}[N] < \infty$.

4. It follows that $\mathbb{P}(N = \infty) = 0$.

Converse? It is NOT true that

$$\mathbb{P}(A_n \text{ i.o.}) = 0 \Longrightarrow \sum \mathbb{P}(A_n) < \infty.$$

Counter-example:

- Let $\mathbb{P}$ be uniform on $[0, 1]$.
- Take $A_n = [0, 1/n]$.
- We have $(A_n \text{ i.o.}) = \{0\}$, hence $\mathbb{P}(A_n \text{ i.o.}) = 0$.
- But $\sum \mathbb{P}(A_n) = \infty$.

But if we add independence, it is true: (BC 2) If

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty,$$

and the events A_n are independent, then $\mathbb{P}(A_n \text{ i.o.}) = 1$.

Example: a monkey on a typewriter will write the complete work of Shakespeare infinitely often. Some notation first:

- Denote the work of Shakespeare by $x_0, x_1, x_2, \ldots, x_{k-1}$ where each x_i is a letter.
- Let M_i denote the letter pressed by the monkey at step i. Assume the $\{M_i\}$ are independent.
- Define $A_n \coloneqq (M_n = x_0, M_{n+1} = x_1, \ldots, M_{n+k-1} = x_{k-1})$.

Problem? A_n are not independent! We will actually prove something stronger. Informally, the idea is that we are going to show the monkey write the work of Shakespeare *and with the first letter written say Jan 1st* infinitely often. Formally:

- Define $A'_n \coloneqq (M_{kn} = x_0, M_{kn+1} = x_1, \ldots, M_{kn+k-1} = x_{k-1})$.
- The A'_n are independent by construction.
- Hence, by BC 2, $\mathbb{P}(A'_n \text{ i.o.}) = 1$.
- Now, $(A'_n \text{ i.o.}) \subset (A_n \text{ i.o.})$, so $\mathbb{P}(A_n \text{ i.o.}) \geq \mathbb{P}(A'_n \text{ i.o.}) = 1$.

6.3 Weak law of large number (WLLN)

Let $X_i : \Omega \to \mathbb{R}$ be iid, with $\mathbb{E}[X_i] = \mu, |\mu| < \infty$, and defined $S_n = X_1 + \cdots + X_n$. To capture our intuition of how repeated random processes behave ("frequencies approach probabilities"), we would like to be able to write something like:

$$\frac{1}{n} S_n \to \mu.$$

That raises the question: what do we mean by "$\to$"? So far, the definition we used was *pointwise limits*, meaning that for all $\omega \in \Omega$,

$$\lim_{n\to\infty} \frac{1}{n} S_n(\omega) = \mu.$$

Note here that the LHS is a random variable, so the right hand side should be interpreted as a constant random variable (i.e. such that $\mu(\omega) = \mu$ for all ω).

It is too much to ask for such convergence to hold in general. Consider for instance the train example, where ω corresponds to an infinite trajectory. Then for the trajectory ω_0 where the train always goes right (s.t. $X_i(\omega_0) = +1$ for all i). Then

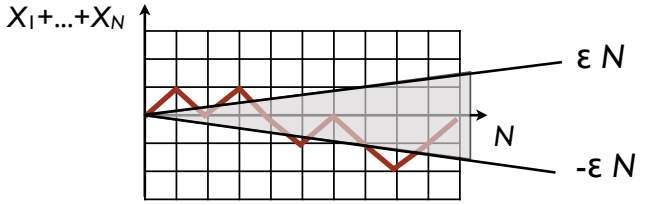
$$\lim \frac{1}{n} S_n(\omega_0) = 1 \neq \mu = 0.$$

Instead: we will relax the notion of convergence. There are several ways to do this, which are useful in different contexts. Let us start with one that gives us a simple proof of the LLN.

Proposition: let $X_i : \Omega \to \mathbb{R}$ be iid, with $\mathbb{E}[X_i] = \mu, |\mu| < \infty$, and defined $S_n = X_1 + \cdots + X_n$. Then for all $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\left| \frac{1}{n} S_n - \mu \right| \geq \epsilon \right) = 0.$$

Visualization: suppose $\mu = 0$, and note that $\left(\left|\frac{1}{n} S_n - 0\right| \geq \epsilon\right) = (|S_n| > \epsilon n)$. On a graph where the x-axis is n and the y-axis is S_n (the same picture we used for train trajectories), consider the cone specified by ϵn and $-\epsilon n$. The event $(|S_n| > \epsilon n)$ can be understood as selecting the trajectories that are outside the cone at step n. The probabilities of these events should go to zero as $n \to \infty$.



Proof: we will start with a proof that uses an extra assumption, $\mathbf{Var}X < \infty$. I will then give a sketch of how to get rid of this condition.

1. First, assume without loss of generality that $\mu = 0$ (otherwise, set $X_i' = X_i - \mu$).

2. Use Chebyshev:

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{1}{n}S_n\right| \geq \epsilon\right) &\leq \frac{\mathbb{E}\left|\frac{1}{n}S_n\right|^2}{\epsilon^2} \\
&= \frac{\mathbb{E}[X_1 + \cdots + X_n]^2}{n^2\epsilon^2} \\
&= \frac{\sum_{i=1}^n \mathbb{E}[X_i^2] + \sum_{(i,j):i\neq j} \mathbb{E}[X_i X_j]}{n^2\epsilon^2} \\
&= \frac{n\mathbb{E}X_i^2}{n^2\epsilon^2} \\
&= \frac{\text{constant}}{n}.
\end{aligned}$$

3. Taking limits on both sides:

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}S_n\right| \geq \epsilon\right) \leq \lim_{n\to\infty} \frac{\text{constant}}{n} = 0.$$

Note: we actually only used pairwise uncorrelation in this proof.

Sketch: for how to lift the finite variance condition. The idea is to first fix $x > 0$, and to write the following decomposition:

$$X_i = \underbrace{X_i \mathbf{1}[|X_i| \leq x]}_{Y_i} + \underbrace{X_i \mathbf{1}[|X_i| > x]}_{Z_i}.$$

Now we can use our previous result on Y_i since a bounded random variable will necessarily have finite variance. As for Z_i, we can get rid of it by letting $x \to \infty$ and using DCT (where we use $|Z_i| \leq |X_i|$ and our assumption that $|\mu| < \infty$ and hence $\mathbb{E}|X_i| < \infty$).

6.4 Convergence in probability

The definition of "$\to$" used in the previous section is used in other contexts, so let us give a name to it:

Definition: a sequence of random variables $Y_i : \Omega \to \mathbb{R}$ converges in probability to a random variable $Y : \Omega \to \mathbb{R}$ if, for all $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|Y_n - Y| > \epsilon) = 0.$$

Notation: $Y_n \xrightarrow{\mathbb{P}} Y$.

6.5 Convergence almost surely

Question: can we come up with other notions of convergence? Yes, we will see many alternative, starting with "almost sure" convergence ("$Y_n \xrightarrow{a.s.} Y$"). Why is this useful? With almost sure convergence, it is harder to prove the LLN, but once this is done, it is easier to establish corrolaries, e.g. that $Y_n \xrightarrow{a.s.} Y$ and $Y'_n \xrightarrow{a.s.} Y'$ implies $Y_n + Y'_n \xrightarrow{a.s.} Y + Y'$.

Definition: a sequence of random variables $Y_i : \Omega \to \mathbb{R}$ converges almost surely (a.s.) to a random variable $Y : \Omega \to \mathbb{R}$ if

$$\mathbb{P}(Y_n \to Y) = 1,$$

where

$$(Y_n \to Y) \coloneqq \{\omega \in \Omega : \lim_{n\to\infty} |Y_n(\omega) - Y(\omega)| \text{ exists and is } = 0\}.$$

Note: with this last notation, another way to write convergence pointwise is $(Y_n \to Y) = \Omega$. From this, it is clear that convergence pointwise implies convergence a.s. but not vice-versa.

Notation: $Y_n \xrightarrow{a.s.} Y$.

Lemma connecting this notion of convergence back to convergence in probability as well as the notions of i.o. and ev.: $X_n \xrightarrow{a.s.} X$ if and only if for all $\epsilon > 0$, $\mathbb{P}(|X_n - X| > \epsilon \text{ i.o.}) = 0$.

Proof: we have:

$$\omega \in (X_n \to X) \iff \forall \epsilon > 0, \omega \in (|X_n - X| \leq \epsilon \text{ ev.}),$$

and hence:

$$\begin{aligned} X_n \xrightarrow{a.s.} X &\iff \forall \epsilon > 0, \mathbb{P}(|X_n - X| \leq \epsilon \text{ ev.}) = 1 \\ &\iff \forall \epsilon > 0, \mathbb{P}(|X_n - X| > \epsilon \text{ i.o.}) = 0. \end{aligned}$$

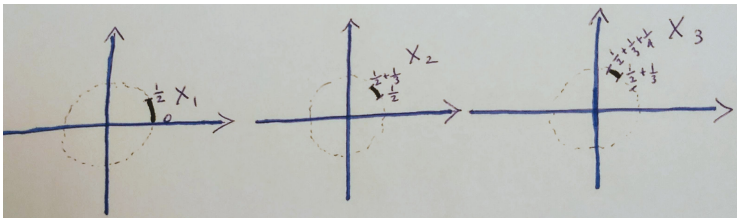
Strong law of large numbers: under the same conditions as the WLLN, this says

$$\frac{1}{n} S_n \xrightarrow{a.s.} \mu.$$

Proposition: convergence in probability does not imply convergence a.s. in general.

Counter-example: moving blip. Let $\Omega =$ points on a circle, $\mathbb{P}$ be uniform on the circle, and X_i be defined as:

$$\begin{aligned} X_1 &= \mathbf{1}_{[0,1/2] \bmod 2\pi} \\ X_2 &= \mathbf{1}_{[1/2,1/2+1/3] \bmod 2\pi} \\ X_3 &= \mathbf{1}_{[1/2+1/3,1/2+1/3+1/4] \bmod 2\pi} \\ &\vdots \end{aligned}$$



Proposition: convergence a.s. implies convergence in probability.

Proof: according to our earlier lemma, we have for all $\epsilon > 0$, $\mathbb{P}(|X_n - X| > \epsilon \text{ i.o.}) = 0$. Then, if we let $A_n \coloneqq (|X_n - X| > \epsilon)$:

$$\begin{aligned} 0 &= \mathbb{P}\left(\bigcap_{k=1}^{\infty} \bigcup_{n \geq k} A_n\right) \\ &= \lim_{k \to \infty} \mathbb{P}\left(\bigcup_{n \geq k} A_n\right) \text{ (by monotonicity of } \mathbb{P}) \\ &\geq \lim_{k \to \infty} \mathbb{P}(A_k) \geq 0. \end{aligned}$$

It follows that $\lim_{k \to \infty} \mathbb{P}(A_k) = 0$.

6.6 Toward Central Limit Theorems

Central limit theorems (CLT) are tools to approximate the distribution of a sum of certain random variables, $S_n = \sum_{i=1}^{n} X_i$. Before we describe CLTs and their proofs, let us motive the problem by first computing the exact distribution of S_n to illustrate why it is non-trivial.

6.7 Exact distribution of sums of random variables

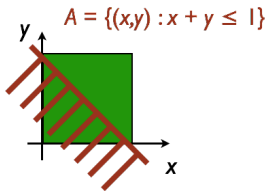
Let us say we have two iid random variables X and Y with densites f_X and f_Y. What is the distribution of $Z = X + Y$? It is not as easy as it looks, for example you should certainly not average the densities (check! either by playing Settlers of Catan, or deriving the pmf of the sum of two dice).

Example: to make things concrete, suppose X and Y are iid uniform on $[0, 1]$.

Compute the CDF of Z first: at a fixed point, say one, $F_Z(1)$. To do so let us use the techniques of Section 5.5:

$$\begin{aligned} \mathbb{P}(Z \leq 1) &= \mathbb{P}(X + Y \leq 1) \\ &= \int \mathbf{1}[(x, y) \in A] f_X(x) f_Y(y) \, \mathrm{d}x \, \mathrm{d}y, \end{aligned}$$

where the region A is shown in red below.



Hence

$$\mathbb{P}(Z \leq 1) = \int_{-\infty}^{\infty} \int_{-\infty}^{1-x} f_X(x) f_Y(y) \, \mathrm{d}y \, \mathrm{d}x = \frac{1}{2}.$$

More generally,

$$\begin{aligned} \mathbb{P}(Z \leq z) &= \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f_X(x) f_Y(y) \, \mathrm{d}y \, \mathrm{d}x \\ &= \int_{-\infty}^{\infty} f_X(x) \int_{-\infty}^{z-x} f_Y(y) \, \mathrm{d}y \, \mathrm{d}x \\ &= \int_{-\infty}^{\infty} f_X(x) F_Y(z-x) \, \mathrm{d}x \end{aligned}$$

Computing the density: by differentiation of the above expression with respect to the argument z.

$$\begin{aligned} f_Z(z) &= \frac{\mathrm{d}F_Z(z)}{\mathrm{d}z} \\ &= \frac{\mathrm{d}}{\mathrm{d}z} \int_{-\infty}^{\infty} f_X(x) F_Y(z-x) \, \mathrm{d}x. \end{aligned}$$

If only we could interchange the order of differentiation and integration that would lead to a nice expression:

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_X(x) \frac{\mathrm{d}}{\mathrm{d}z} F_Y(z-x) \, \mathrm{d}x \\ &= \int_{-\infty}^{\infty} f_X(x) f_Y(z-x) \, \mathrm{d}x. \end{aligned}$$

In the next section, we justify this swap.

Exercise: show that if you sum two uniform, you get the following density, called the triangular density:

Summing more than two variables. In theory, generalizing this argument to more than two variables is simple, just iterate the above argument. In practice,

for summing n random variables we end up with a $n-1$ dimensional integral to compute. It can be tricky to compute the integral exactly, motivating the need for CLTs.[7]

6.8 Interlude: exchanging the order of integration and differentiation

Here we present some useful theorem justifying swapping the order of differentiation and integration. Doing so will not only justify the swap in last section, but will also be needed soon in our CLT proof.

Example: "reparameterization trick," a technique very popular in machine learning. The idea is that you have a collection of probability models indexed by a parameter θ (as in Section 2.6 for example) and you seek to compute

$$\nabla \mathbb{E}_\theta[h(X)].$$

Here the difficulty is that the distribution of X depends on θ. The idea with the reparameterization trick is to remove this dependency by writing $X = g(Z, \theta)$, where the distribution of Z does not depend on θ. From Section 2.26, we know how to do this: set Z to be uniform, and $g(\cdot, \theta)$ to the inverse CDF of $F_\theta(\cdot)$. After doing this, the hope is to swap the order of differentiation and integration:

$$\begin{aligned} \nabla \mathbb{E}_\theta[h(X)] &= \nabla \mathbb{E}[h(g(Z,\theta))] \quad \text{(always possible)} \\ &= \mathbb{E}[\nabla h(g(Z,\theta))] \quad \text{(have to be careful here).} \end{aligned}$$

If we can do this, that would be nice, since we can then use the Law of large number to approximate the right hand side:

$$\mathbb{E}[\nabla h(g(Z,\theta))] \approx \frac{1}{N} \sum_{n=1}^{N} \nabla h(g(Z^{(n)},\theta)),$$

where $Z^{(n)}$ are iid samples, and the inner gradient is often computed using Automatic Differentiation techniques.

Counter-example. Let us see a concrete example where the above trick *does not* work, to emphasize the importance of checking conditions for integral-differential swaps. Let us say $X \sim \text{Bern}(\theta)$. We can reparametrize with $Z \sim \text{Unif}(0,1)$ and $g(z,\theta) = \mathbf{1}[z < \theta]$. Indeed, $g(Z,\theta) \sim \text{Bern}(\theta)$. We have:

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \underbrace{\int_0^1 g(z,\theta)\,\mathrm{d}z}_{=\theta} = 1,$$

[7] But there is one clever trick that exploit the special structure of these iterated integrals, called the Fast Fourier Transform. It uses similar techniques as the CLT. But it is still more expensive than the CLT and for sums of random vectors it can be prohibitive.

but on the other hand:

$$\int_0^1 \underbrace{\frac{\mathrm{d}}{\mathrm{d}\theta} g(z,\theta)}_{=0 \text{ } almost \text{ everywhere}} \mathrm{d}z = 0.$$

Theorem for swaps: suppose

1. $g(\cdot, \theta)$ is integrable for all $\theta \in [a, b]$,
2. $g(x, \theta)$ is differentiable for all x and θ,
3. there is an integrable envelope $h(x)$ such that $|\,\mathrm{d}g(x,\theta)/\,\mathrm{d}\theta| \leq h(x)$ for all x and θ,

then both sides of the equation below are well defined and equal:

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \int g(x,\theta)\mu(\,\mathrm{d}x) = \int \frac{\mathrm{d}}{\mathrm{d}\theta} g(x,\theta)\mu(\,\mathrm{d}x).$$

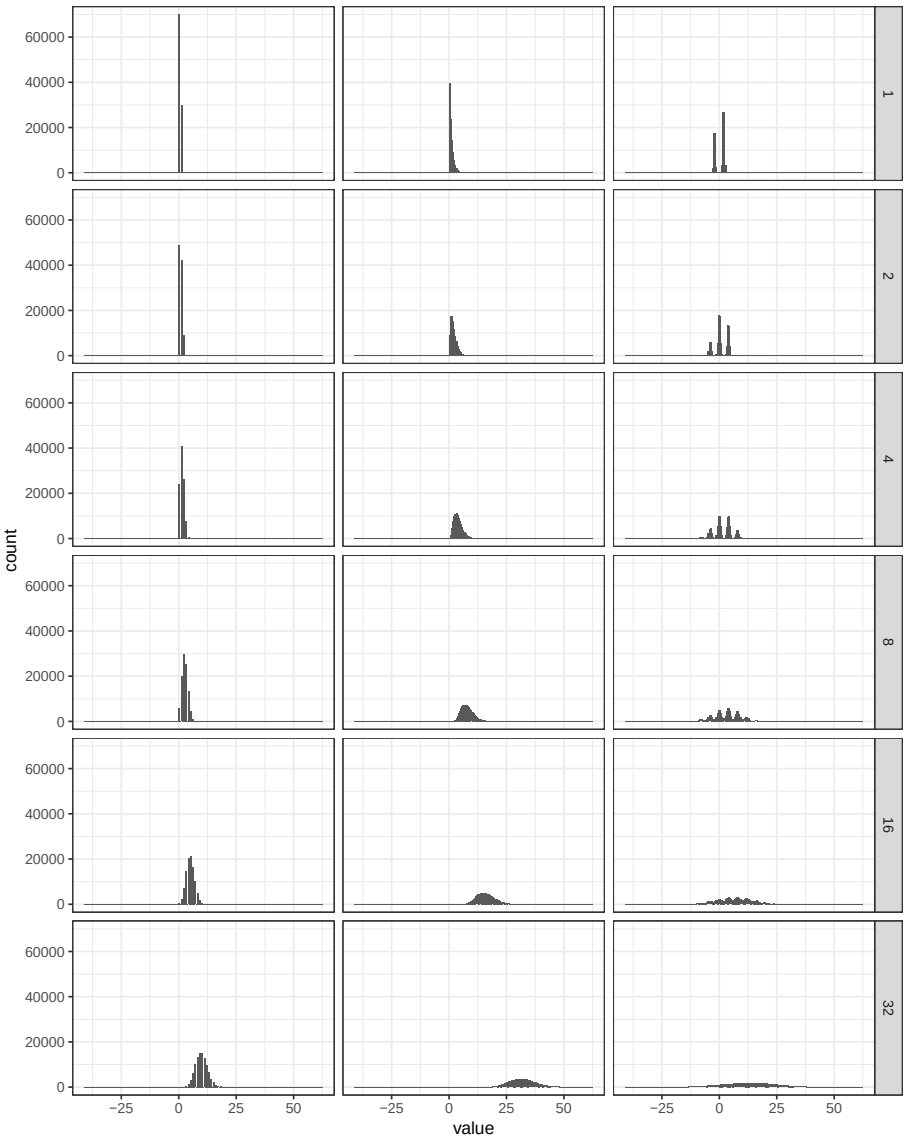
Exercise: find the condition that does not hold in the counter example.

Exercise: use DCT and the mean value theorem to prove the theorem for swaps.

6.9 CLT: numerical exploration and intuition

We have seen how computing the distribution of sums $S_n = X_1 + \cdots + X_n$ is tedious, even for iid random variables. Now, as promised in the introduction of this chapter, the good news is that the distribution of such sums typically becomes more and more regular.

Numerical example. To start with, let us look, for 3 possible distributions for X_i, what the distribution of the sums look like when we increase n. The columns are three different distributions for X_1, and each row shows S_n for $n \in \{1, 2, 4, 8, 16, 32\}$.



As you can see, we start with three very different distributions (first row, $n = 1$, shows the distribution of $S_1 = X_1$):

1. A discrete distribution: Bernoulli distribution $X_i \sim \text{Bern}(0.3)$.
2. A continuous distribution with support on $[0, \infty)$: an exponential distribution, $X_i \sim \text{Exp}(1)$
3. A multimodal distribution: with a density given by a *mixture* of normal distribution $f_{X_i}(x) = 0.4 f_{\text{N}}(x; -2, 0.1) + 0.6 f_{\text{N}}(x; 2, 0.1)$ where $f_{\text{N}}(\cdot; \mu, \sigma^2)$ is the normal density (defined below).

Normal approximation: notice that for $n = 32$, all three look the same, up to stretching and translation. In fact, all three look like normal distributions! This is surprising, since the normal is continuous on $(-\infty, \infty)$ and unimodal, which neither of the three starting distribution X_1 remotely looked like! Our next big goal is to understand why!

Heuristic. Before going into proofs, however, the above numerical study already motivates the following heuristic:

1. Find the mean μ and variance σ^2 of S_n. (why? in the "normal approximation" paragraph above, we say all three "look like normal distributions," now we need to find which normal distribution!
2. Let Z_n denote a normal distribution with mean μ and variance σ^2, defined as the random variable with density
$$f_{\mathrm{N}}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$
3. Use the approximation $S_n \overset{d}{\approx} Z_n$.

Interpretation of approximation: $S_n \overset{d}{\approx} Z_n$? We will need more theoretical understanding to refine what this means exactly. For now think about it as:

1. $\mathbb{P}(S_n \leq x) \approx \mathbb{P}(Z_n \leq x)$, or
2. $\mathbb{E}[g(S_n)] \approx \mathbb{E}[g(Z_n)]$ for nice g, e.g. bounded continuous.

CLTs and leaps of faith. CLTs attempt to justify the above heuristic, either by telling us that the approximation will become arbitrarily accurate as $n \to \infty$ (basic CLT), or by providing explicit bounds on the error (Berry-Esseen theorem), or by relaxing the conditions on the random variables. In some situations, the theory might not bridge completely to your problem, but even in such cases, it might be still useful to take a "leap of faith" and use the normal approximation (in particular, in a situation where you might not be able to give an answer at all otherwise).

Exercise: You seed a Petri dish with a colony of 1e9 bacteria. Each day, the number of bacteria either:

- stays the same, with probability 2/5;
- doubles, with probability 3/10;
- quadruples, with probability 1/5;
- is divided by a factor of two, with probability 1/10.

Ten days later, what is the probability that there are more than 30e9 bacteria?

Hint: write number of bacteria as $2^{\sum X_i}$, and rewrite the probability you want to obtain into one that can be computed using the normal approximation.

6.10 Basic CLT

Assumptions: the random variables we sum are independent and identically distributed and have finite variance. Also, without loss of generality assume $\mathbb{E}X_i = 0$ (otherwise, center the random variables).

Relax: there are CLTs for non-identical random variables, also for random vectors, and finally, for dependent random variables (e.g. certain types of Markov chains).

In relation with LLN: sure, the LLN gives us $\frac{1}{n}S_n \xrightarrow{a.s.} 0$. But this does not give us much information about how to approximate the distribution of S_n for finite n.

Exercise: show that in fact, $\frac{1}{n^\alpha}S_n \xrightarrow{\mathbb{P}} 0$ for all $\alpha > 1/2$ (the same can be done almost sure with a bit more work).

Idea: the above exercise suggest that $\alpha = 1/2$ is a critical point. A related heuristic is that for a symmetric simple random walk, the expected distance from the origin at step n is roughly $\sqrt{n}$.

Theorem: indeed, if $X_1, X_2, \ldots$ are idd with $EX_1^2 = 1$ and $\mathbb{E}X_1 = 0$, then for any $x \in \mathbb{R}$:

$$\mathbb{P}\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}} \leq x\right) \to \mathbb{P}(Y \leq x), \text{ as } n \to \infty,$$

where Y is a standard normal.

6.11 Exercise set 4

1. First problem in Section 6.8.
2. Second problem in Section 6.8 again.
3. One problem in Section 6.9.
4. One problem in Section 6.10.

6.12 Solutions for exercise set 4

Question 1: there is a discontinuity at $\theta = x$, so we cannot say the function is differentiable at all points, which is required for this theorem (in contrast to many other measure theoretic ones, where almost everywhere is enough, here we really need everywhere).

Question 2: the first idea is to transform the statement into one about sequences of function in the optic of applying DCT. Let $\theta_n \to \theta$ be arbitrary in $[a, b]$. Let

$$h_n(x, \theta) = \frac{g(x, \theta_n) - g(x, \theta)}{\theta_n - \theta}.$$

Note that

$$\lim_{n\to\infty} h_n(x,\theta) = \frac{\mathrm{d}g(x,\theta)}{\mathrm{d}\theta}.$$

Next let us bound $|h_n|$ by h:

$$|h_n(x,\theta)| \leq \sup_{\theta\in[a,b]} \left|\frac{\mathrm{d}g(x,\theta)}{\mathrm{d}\theta}\right| \leq h(x),$$

where we use the mean value theorem in the first inequality and the assumption on h in the theorem statement for the second inequality. With this bound and the second assumption on h that it is an integrable envelope, we can apply DCT and obtain the result.

Question 3: Following the hint in the question we will seek to express the number of bacteria at day 10 in terms of $2^{\sum_{i=1}^{10} X_i}$. To see this, consider the possible day-over-day changes in the number of bacteria in the Petri dish: (i) stays the same; (ii) doubles; (iii) quadruples; (iv) or is halved. Each of the changes in the number can be represented as a power of two. For example, if the number starts with value 1e9 then doubles, stays the same, quadruples, and halves in the subsequent 4 days, the number of bacteria can be expressed succinctly as:

$$\text{the number of bacteria after 4 days} = 1e9 \times 2^1 \times 2^0 \times 2^2 \times 2^{-1} = 1e9 \times 2^{1+0+2-1} = 1e9 \times 2^2 = 4e9.$$

To generalize the number of bacteria over all possible paths, we define the random variable X_i to correspond to the power of 2 that the number "grows" in each day. Therefore, over a ten day period the (random) value of the number of bacteria is represented as $2^{\sum_{i=1}^{10} X_i}$. Since we want the number of bacteria to be more than 30 times of its initial value, we need to calculate:

$$\mathbb{P}(\text{more than 30e9}) = \mathbb{P}\left(2^{\sum_{i=1}^{10} X_i} \geq 30\right) = \mathbb{P}\left(\sum_{i=1}^{10} X_i \geq \log_2(30)\right).$$

However, we do not know the distribution of $\sum_{i=1}^{10} X_i$! Notwithstanding this limitation, this quantity is still possible to approximate. We will use the Central Limit theorem to provide an approximation which is easy to calculate. First, let $S_n := \sum_{i=0}^{n} X_i$, then:

$$S_{10} \overset{approx.}{\approx} a + bZ$$

where Z follows the standard normal distribution, with:

$$a = n\mu = n \cdot \mathbb{E}(X_i) = 10\left(\frac{2}{5}\cdot 0 + \frac{3}{10}\cdot 1 + \frac{1}{5}\cdot 2 + \frac{1}{10}\cdot(-1)\right) = 10 \cdot 0.6 = 6$$

$$b = \sigma\sqrt{n} = \sqrt{\mathbf{Var}(X_i)\cdot n} = \sqrt{(\mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2)\cdot 10} = \sqrt{(1.2 - 0.6^2)\cdot 10} = \sqrt{\frac{42}{5}} \approx 2.898,$$

where the fact that $\mathbb{E}(X_i^2) = \frac{2}{5}\cdot 0^2 + \frac{3}{10}\cdot 1^2 + \frac{1}{5}\cdot 2^2 + \frac{1}{10}\cdot(-1)^2 = 1.2$ was used.

Therefore,

$$\begin{aligned}\mathbb{P}(\text{more than 30e9}) &= \mathbb{P}\left(2^{\sum_{i=1}^{10} X_i} \geq 30\right) = \mathbb{P}\left(\sum_{i=1}^{10} X_i \geq \log_2(30)\right) \\ &= \mathbb{P}(S_{10} \geq \log_2 30) = P\left(Z \geq \frac{\log_2 30 - 6}{\sqrt{\frac{42}{5}}}\right) \\ &= 1 - \Phi\left(-0.3772\right) \approx 0.647.\end{aligned}$$

Question 4:

1. First, assume without loss of generality that $\mu = 0$ (otherwise, set $X_i' = X_i - \mu$).

2. Use Chebyshev:

$$\begin{aligned}\mathbb{P}\left(\left|\frac{1}{n^\alpha} S_n\right| \geq \epsilon\right) &\leq \frac{\mathbb{E}\left|\frac{1}{n^\alpha} S_n\right|^2}{\epsilon^2} \\ &= \frac{\mathbb{E}[X_1 + \cdots + X_n]^2}{n^{2\alpha}\epsilon^2} \\ &= \frac{\sum_{i=1}^n \mathbb{E}[X_i^2] + \sum_{(i,j):i\neq j} \mathbb{E}[X_i X_j]}{n^{2\alpha}\epsilon^2} \\ &= \frac{n\mathbb{E}X_i^2}{n^{2\alpha}\epsilon^2} \\ &= \frac{\text{constant}}{n^{2\alpha-1}}.\end{aligned}$$

3. Taking limits on both sides:

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n^\alpha} S_n\right| \geq \epsilon\right) \leq \lim_{n\to\infty} \frac{\text{constant}}{n^{2\alpha-1}} = 0.$$

   provided $2\alpha - 1 > 0$, i.e. $\alpha > 1/2$.

6.13 Types of convergence: big picture

We would like to generalize the notion of convergence used in the CLT. We would like this new notion of convergence, called convergence in distribution, to sit nicely in our hierarchy of convergence modes as follows.

Covered so far:

$$X_n \to X \Longrightarrow X_n \xrightarrow{a.s.} X \Longrightarrow X_n \xrightarrow{\mathbb{P}} X.$$

Next:

$$X_n \xrightarrow{\mathbb{P}} X \Longrightarrow X_n \xrightarrow{d} X.$$

6.14 Weak convergence

Motivation: so far, the two types of convergence we have covered assume that all random variables and the limit live in the same space: $X_i : \Omega \to \mathbb{R}$, $X : \Omega \to \mathbb{R}$. This could create problem when we formalize the central limit theorem. Why?

Solution: define convergence with respect to objects that get rid of Ω while characterizing the distribution of the random variables.

Definition 1: we say $X_n : \Omega_n \to \mathbb{R}$ converges in distribution to $X : \Omega \to \mathbb{R}$, denoted $X_n \xrightarrow{d} X$, if for all bounded continuous function $g : \mathbb{R} \to \mathbb{R}$ (called "test function"), we have

$$\lim_{n\to\infty} \int g \, \mathrm{d}\mathbb{P}_{X_n} = \int g \, \mathrm{d}\mathbb{P}_X.$$

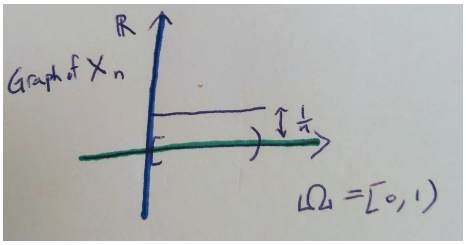
Note: more generally, weak convergence is really about convergence of measures. Similarly to the above definition, the notation $\mu_n \xrightarrow{d} \mu$ means that for all bounded continuous function g,

$$\lim_{n\to\infty} \int g \, \mathrm{d}\mu_n = \int g\mu.$$

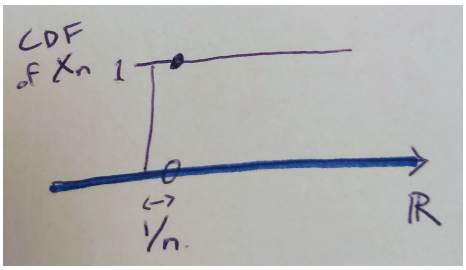
Definition 2: we say $X_n : \Omega_n \to \mathbb{R}$ converges in distribution to $X : \Omega \to \mathbb{R}$, denoted $X_n \xrightarrow{d} X$, if for all x with $\mathbb{P}(X = x) = 0$, $F_{X_n}(x) \to F(x)$. In other words, convergence of the CDFs at points that are not atoms under X.

Proposition (Portmanteau): definition 1 and 2 are in fact equivalent. In fact, search Portmanteau in your favourite reference, there are many more equivalent definitions, e.g. a useful variant is to replace "bounded continuous" by "Lipschitz" (i.e. such that there is a constant $K > 0$ such that $|g(x_1) - g(x_2)| \leq K|x_1 - x_2|$.

Intuition regarding why we do not require convergence of the CDF at atomic points: otherwise, our nice hierarchy in Section 6.13 would not hold! Consider: $X_n = 1/n$. Show it converge in probability to zero:



however the limit of the cdfs at zero is equal to zero but the cdf of the limit at zero is one:



We can now write the CLT as follows: if $X_1, X_2, \ldots$ are idd with $EX_1^2 = 1$ and $\mathbb{E}X_1 = 0$, then:

$$\frac{X_1 + \cdots + X_n}{\sqrt{n}} \xrightarrow{d} Y,$$

where Y is a standard normal.

Note: this statement is not true in probability or a.s. (this can be proven using 'Kolmogorov zero-one law').

6.15 Overview of some properties of convergence of r.v.'s

Scheffé's theorem: if X_n have density f_n such that $f_n \to f$ pointwise, where f is the density of X, then $X_n \xrightarrow{d} X$.

Proof: dominated convergence.

Continuous mapping I: if g is continuous, $X_n \xrightarrow{a.s.} X$, then $g(X_n) \xrightarrow{a.s.} g(X)$.

Proof: follows directly from definition of continuity.

Continuous mapping II: if g is continuous, $X_n \xrightarrow{d} X$, $g(X_n) \xrightarrow{d} g(X)$.

Proof: trivial, provided you use the right equivalent definition provided by Portmanteau.

Continuous mapping III: if g is continuous, $X_n \xrightarrow{\mathbb{P}} X$, then $g(X_n) \xrightarrow{\mathbb{P}} g(X)$.

Note: this can be relaxed to accommodate a set of discontinuity D as long as it has probability zero under the asymptotic distribution.

Template for many results: "$X_n \xrightarrow{\square} X$ and $Y_n \xrightarrow{\square} Y$ implies $X_n \circ Y_n \xrightarrow{\square} X \circ Y$" where you should replace $\square$ and $\circ$ by any of these combinations (note that some combinations require extra assumptions):

1. $\square$ =a.s. and $\circ = +$, or $-, *, /$.
2. $\square$ =p and $\circ = +$, or $-, *, /$.
3. $\square$ =d and $\circ = +$, AND X_n, Y_n, X, Y independent.
4. $\square$ =d and $\circ = +$, or $-, *, /$, AND Y = constant ("Slutsky's theorem").

Example: of why we have to be careful, especially with convergence in distribution. Let Y, X, X_n all be iid N(0,1), and $Y_n = -X_n$. Then $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$, but $X_n + Y_n \xrightarrow{d} 0 \overset{d}{\neq} X + Y$. Here we cannot use Slutsky's theorem because neither X_n nor Y_n converge to a constant.

Main tools: used to proved these results.

- Subsequence characterization of convergence in probability: $X_n \xrightarrow{\mathbb{P}} X$ if and only if for all subsequence n_k, there is a further subsequence n_{k_i} such that $X_{n_{k_i}} \xrightarrow{a.s.} X$.

- Skorokhod representation: suppose $X_n \xrightarrow{d} X$, then there exists Y_n and Y such that $Y_n, Y : \Omega \to \mathbb{R}$, $Y_n \overset{d}{=} X_n$, $Y \overset{d}{=} X$, and $Y_n \xrightarrow{a.s.} Y$.

6.16 Towards a proof of the CLT: generating functions

To prove the CLT, we first need to look at tools that will help us study sums of random variable. These tools are called generating functions. We go over generating functions because they are useful in many contexts, from theory to computation.

Why generating functions? A generating function associates a function $G_X(s)$ to a random variable X. They are designed with two goals in mind:

1. They *characterize* certain classes of distributions. I.e. just as CDFs, if generating functions are equal (as functions), $G_X = G_Y$, then the corresponding random variables are equal in distribution, $X \overset{d}{=} Y$.

2. The generating function of sums of independent random variables is the product of the individual generating functions: $G_{X+Y}(s) = G_X(s)G_Y(s)$.

Specific applications of generating functions:

1. Most CLT proofs.

2. The Fast Fourier Transform, one of the most important algorithms out there, is heavily based on generating functions. As an example of application of the Fast Fourier Transform, say you want to get the coefficient of the polynomial product $(a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0) \times (b_n x^n + b_{n-1}x^{n-1} + \cdots + b_0)$. How long does it take to compute the coefficients of the product, $c_{2n}x^{2n} + c_{2n-1}x^{2n-1} + \cdots + sc_0$? Naively, $O(n^2)$. Surprisingly, there exists a $O(n \log n)$ algorithm based on the Fast Fourier Transform.[8]

Types of generating function that we will survey:

[8] Many references available online, e.g. `http://www.cs.uleth.ca/~benkoczi/files/fourier-excerpt.pdf`

Probability generating functions: simplest to understand but only characterizes natural number valued random variables. We will start there for pedagogical reasons.

Moment generating functions: which we already encountered when we talked about Chernoff bound in Section 4.5. They are defined for some random variables beyond natural number valued ones, but not all.

Characteristic function: which are always defined but will require a quick review on complex numbers. After covering those we will be ready for the proof of CLT.

6.17 Probability generating functions

Restriction: we only consider natural number valued random variables, i.e. $X : \Omega \to \{0, 1, 2, \dots\}$.

Definition of a probability generating function (PGF):

$$G_X(s) = \mathbb{E}[s^X],$$

which in the context of a natural number valued random variables,

$$G_X(s) = \sum_{i=0}^{\infty} \mathbb{P}(X = i) s^i.$$

Note that this sum might diverge for some values of s. For reason that will become clear soon, we elect to only define it when $\mathbb{E}|s^X| < \infty$ (the terminology is "define G for s where the series is absolutely convergent").[9]

Exercise: compute it for a Bernoulli. Then for a Poisson, with pmf for parameter $\lambda > 0$ given by

$$p(n) = e^{-\lambda} \frac{\lambda^n}{n!} \mathbf{1}[n \in \mathbb{N}].$$

Solution for Bernoulli: if $X \sim \text{Bern}(\theta)$, then $G_X(s) = (1 - \theta) + \theta s$.

Solution for Poisson: if $Y \sim \text{Poi}(\lambda)$, the $G_Y(s) = e^{\lambda(s-1)}$.

Next: do probability generating functions indeed satisfy the two desiderata described in the last section (characterization and nice behaviour for sums)?

First desiderata: nice behaviour with sums. If X and Y are independent, the PGF of their sum is the product of the PGFs, $G_{X+Y} = G_X G_Y$.

Proof: by Tonelli:

$$\mathbb{E}|s^{X+Y}| = \mathbb{E}|s^X s^Y| = \mathbb{E}|s^X| \mathbb{E}|s^Y|,$$

[9] From real analysis, this is is true in a radius of at least one around the origin.

and the right hand side is finite by the absolute convergence assumption introduced in the definition. Then this mean we can apply Fubini and obtain:

$$G_{X+Y}(s) = \mathbb{E}[s^{X+Y}] = \mathbb{E}[s^X]\mathbb{E}[s^Y] = G_X(s)G_Y(s).$$

Exercise: find the PMF of a sum of two independent (1) Bernoullis, (2) Poisson random variables with rates λ_1 and λ_2.

Solution for Bernoulli: $G_{X_1+X_2}(s) = ((1-\theta)+\theta s)^2$.

Solution for Poisson: $G_{Y_1+Y_2}(s) = e^{(\lambda_1+\lambda_2)(s-1)}$.

Observation: closure under sums. Notice that for the Poisson example above, the PGF of $Y_1 + Y_2$ coincides with the PGF of a Poisson with mean parameter $\lambda = \lambda_1 + \lambda_2$. This is not always the case: for example, for the Bernoulli example, we do not have this property. In the Poisson case, can we conclude right away that $Y_1 + Y_2$ is Poisson distributed?

Answer: not yet! A priori, we do not know that some other distribution could yield the same PGF! But we will next argue that in fact this cannot happen!

Second desiderata: characterization. How to show that some other distribution cannot in fact yield the same PGF? By solving the following problem: give me a PGF G, and reconstruct the PMF, $p_k = \mathbb{P}(X = k)$. How to do this? Hint: view $G(s)$ as a polynomial. Start with reconstructing $p_0 = p(0)$.

Solution:

$$G(s) = p_0 + p_1 s + p_2 s^2 + p_3 s^3 + \dots.$$

Hence: $G(0) = p_0$. Next, how to get p_1?

Trick: If you know G you can differentiate it! Get:

$$G'(s) = p_1 + 2p_2 s + 3p_3 s^2 + \dots.$$

Hence: $G'(0) = p_1$.

More generally:

$$p_n = \frac{G^{(n)(0)}}{n!}.$$

Application (reading): branching processes. If a cell has a number of offspring that is Poisson-distributed with rate λ, and all its descendants recursively also have each iid Poisson-distributed descendants, what is the probability p that the population eventually dies off? If $\lambda < 1$, $p = 1$! For $\lambda \geq 0$, we can have $p < 1$, and even compute it from the PGF! Please read Section 5.4 in Grimmett and Stirzaker.

Application (optional): probability generating functions are important for the analysis of time series for count data, see for example [5].

6.18 Moment generating function

How to extend the ideas of last section to more general random variables? First try: moment generating functions (MGF).[10]

Idea: in the previous section, we looked at the expectation of s^X. If X is non integer-valued, and we plug in a negative s, we can get $\sqrt{-1}$. In an attempt to avoid complex numbers, the MFG $M(t)$ use the reparameterization $s = e^t > 0$ to avoid complex numbers, yielding:[11]

$$M(t) = \mathbb{E}[e^{tX}].$$

As with the PGF, this expectation may diverge for some values of t. As we shall see this is more problematic here since we get into situations where $M(t)$ is defined only at $t = 0$, in which case the MGF clearly does not characterizes the distribution.

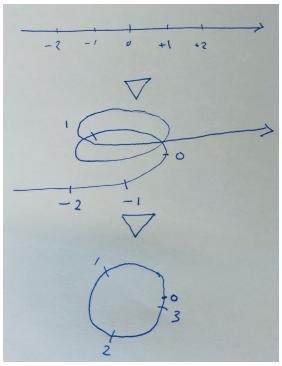
Exercise: compute the MGF of a standard normal distribution. Hint: complete the square to use known normalization of the normal density.

Answer: $\varphi(t) = \exp(t^2/2)$.

6.19 Characteristic function

We would like to have a construction that is defined for all real random variables, and that has our two desiderata (nice behavior under independent sums, characterization).

Idea: in the last section we use the function $f(t) = e^t$, which is a function into an unbounded space, $f : \mathbb{R} \to [0, \infty)$. Let us try to map into a bounded space. To do this, let us wrap around the line into a circle as follows:



Let us formalize this construction using some light complex analysis.

[10] Also known as the *Laplace transform*, up to reparameterization.

[11] However, as we will see soon, complex numbers are in some sense inevitable to do this in full generality.

Recall: a complex number $c \in \mathbb{C}$ can be written as: $c = a + bi$ where $i = \sqrt{-1}$, i.e. such that $i^2 = -1$. The *modulus* of a complex number is defined as $|c| = \sqrt{(a^2 + b^2)}$.

Definition: a complex random variable is defined as $Z = X + iY$, and we define its expectation as $\mathbb{E}Z = \mathbb{E}X + i\mathbb{E}Y$.

Some nice tricks with Taylor series: let us say x is real.

$$\begin{aligned} e^{ix} &= 1 + (ix) + \frac{(ix)^2}{2!} + \frac{(ix)^3}{3!} + \dots \\ &= 1 + (ix) - \frac{x^2}{2!} - \frac{i(x)^3}{3!} + \dots \\ &= (1 - \frac{x^2}{2!} + \dots) + i(x - \frac{x^3}{3!} + \dots) \\ &= \cos x + i \sin x. \end{aligned}$$

So, $|e^{ix}| = 1$ and we have something like we wanted in our picture above! Moreover $e^{i(x+y)} = e^{ix}e^{iy}$ (check!). Now clearly, if I give you some real value random variable Y,

$$\mathbb{E}[e^{itY}] = \mathbb{E}[\cos(tY)] + i\mathbb{E}[\sin(tY)],$$

is always well defined and finite since sin and cos are bounded!

Success! Define the characteristic function as $\varphi_Y(t) = \mathbb{E}[e^{itY}]$. It satisfies both our desiderata!

Proof: See Durrett. Based on the following idea: if I give you φ, you can reconstruct Y via the formula

$$\mathbb{P}(Y \in (a, b]) = \lim_{T \to \infty} \int_{-T}^{T} \frac{e^{-ita} - e^{itb}}{it} \varphi(t) \, \mathrm{d}t.$$

6.20 Further properties of characteristic functions

Property P0: $X_n \xrightarrow{d} X \iff \varphi_{X_n}(t) \to \varphi_X(t)$.

Property P1: $\varphi_X(0) = 1$

Property P2: $|\varphi_X(t)| \leq 1$

Proof: complex version of Jensen. If g is convex, then $g(\mathbb{E}X) \leq \mathbb{E}[g(X)]$. Take $g(x) = |x|$.

Property P3: $\varphi_X(t)$ is always continuous (viewed as a function of t).

Proof: take $t_n \to t$.

$$\begin{aligned} &\implies it_nX \to itX \text{ a.s.}, \\ &\implies \exp(it_nX) \to \exp(itX) \text{ a.s.}, \end{aligned}$$

Where the last implication follows by continuity of exp. Finally,

$$\varphi_X(t_n) = \mathbb{E}[e^{it_n X}] \xrightarrow{a.s.} \mathbb{E}[e^{itX}] = \varphi_X(t),$$

by using DCT with $|e^{it_n X}| = 1$.

Property P4: If $\mathbb{E}|X| < \infty$, then $\varphi_X(t)$ has a derivative and

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\varphi_X(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}[e^{itX}] \\ &= \mathbb{E}[\frac{\mathrm{d}}{\mathrm{d}t}e^{itX}] \text{ (will use our earlier swap result)} \\ &= \mathbb{E}[iXe^{itX}],\end{aligned}$$

and in particular,

$$\mathbb{E}X = \frac{\varphi_X(0)}{i}.$$

Proof: we use the result in Section 6.8, which we copy here for convenience.

Theorem for swaps: suppose

1. $g(\cdot, \theta)$ is integrable for all $\theta \in [a, b]$,
2. $g(x, \theta)$ is differentiable for all x and θ,
3. there is an integrable envelope $h(x)$ such that $|\,\mathrm{d}g(x,\theta)/\,\mathrm{d}\theta| \le h(x)$ for all x and θ,

then both sides of the equation below are well defined and equal:

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\int g(x,\theta)\mu(\,\mathrm{d}x) = \int \frac{\mathrm{d}}{\mathrm{d}\theta}g(x,\theta)\mu(\,\mathrm{d}x).$$

Back to proof: we use our earlier result with $x = \omega$, $\theta = t$,

$$g(\omega, t) = e^{itX(\omega)},$$

and envelope

$$h(\omega) = |X(\omega)|.$$

For condition 1, we have

$$\int |g(\omega,t)|\mathbb{P}(\,\mathrm{d}\omega) = \int 1\mathbb{P}(\,\mathrm{d}\omega) = 1.$$

For condition 2, we have indeed that $e^{itx} = \cos(tx) + i\sin(tx)$ is continuous for all x and t. For condition 3, we have $\mathbb{E}|X| < \infty$ by assumption.

Property P5: more generally, if $\mathbb{E}[|X^k|] < \infty$,

$$\varphi^{(k)}(t) = \mathbb{E}[(iX)^k e^{itX}].$$

As a consequence, we get the following important special case: if $\mathbb{E}X^2 < \infty$, then by Taylor's Theorem (Peano form) applied around zero, there exists a function h such that $h(t) \to 0$ as $t \to 0$ with:

$$\varphi(t) = 1 + it\mathbb{E}X - \frac{1}{2}t^2\mathbb{E}[X^2] + h(t)t^2.$$

Factoid: the characteristic function of a standard normal distribution is

$$\varphi(t) = e^{-\frac{1}{2}t^2}.$$

This will be the last ingredient needed for proving the CLT. This from should not be a surprise given the closely related form for the MGF. However proving this is surprisingly tricky (requires contour integration methods from complex analysis).

6.21 Proof of basic CLT

Theorem: if $X_1, X_2, \ldots$ are idd with $EX_1^2 = 1$ and $\mathbb{E}X_1 = 0$, then:

$$\frac{X_1 + \cdots + X_n}{\sqrt{n}} \xrightarrow{d} Y,$$

where Y is a standard normal.

Proof: let us fix an arbitrary t, and start by looking at only one variable X_1. We have, by P5, that

$$\varphi_{X_1}(t) = 1 - \frac{1}{2}t^2 + h(t)t^2,$$

where all we know about h is that $h(t) \to 0$ when $t \to 0$.

Trick: to look at a normalized sum, note

$$\begin{aligned}\varphi_{aX}(t) &= \mathbb{E}[e^{it(aX)}] \\ &= \mathbb{E}[e^{i(at)X}] \\ &= \varphi_X(at).\end{aligned}$$

Apply trick to proof: to get the characteristic function of $\frac{S_n}{\sqrt{n}} = \frac{X_1+\cdots+X_n}{\sqrt{n}}$,

$$\begin{aligned}\varphi_{S_n/\sqrt{n}}(t) &= (\varphi_{X_1}(t/\sqrt{n}))^n \\ &= \left(1 - \frac{1}{2}(t/\sqrt{n})^2 + h(t/\sqrt{n})(t/\sqrt{n})^2\right)^n \\ &= \left(1 - \frac{(-t^2/2) + h(t/\sqrt{n})t^2}{n}\right)^n\end{aligned}$$

From real analysis: $e^x = \lim_{n\to\infty}(1 + x/n)^n$. This formula also holds for complex x.

Extension: if $x_n \to x$, then $e^x = \lim_{n\to\infty}(1 + x_n/n)^n$.

Applying this to our setup:

$$x = -t^2/2$$
$$x_n = -t^2/2 + h(t/\sqrt{n})t^2.$$

we get $\lim_{n\to\infty} \varphi_{S_n/\sqrt{n}}(t) = e^{-t^2/2}$.

Conclusion: by P0, it follows that

$$S_n/\sqrt{n} \xrightarrow{d} Z,$$

where $\varphi_Z(t) = e^{-t^2/2}$. By our characterization result and the factoid from last section, it follows that Z is a standard normal random variable.

6.22 CLT: multivariate version

Setup: suppose now we have a sum of iid copies of a d-dimensional mean zero random *vector*, $X_1 = (X_{1,1}, X_{1,2}, \ldots, X_{1,d})$. The interesting bit is that we do *not* assume independence across the d dimensions.

Goal: We are interested in approximating the distribution of the sum of vectors $\sum_{i=1}^n X_i$, where $X_i = (X_{1,i}, X_{i,2}, \ldots, X_{i,d})$. We assume $X_i \stackrel{d}{=} X_j$ for all i, j and that the X_i are independent (across $i \in \{1, \ldots, n\}$, but still allowing for dependence within one vector across dimensions $k \in \{1, \ldots, d\}$).

Assumption: we do need to assume some kind of second moment assumption. In the vector context, this becomes $\mathbb{E}|X_{1,k}X_{1,k'}| < \infty$. This reduces to our usual condition if $d = 1$, so no surprise here.

Theorem: under the above assumptions,

$$\frac{X_1 + X_2 + \cdots + X_n}{\sqrt{n}} \xrightarrow{d} Z,$$

but this raises a few questions: (1) what does convergence in distribution mean for random vectors? (2) what is the limiting random vector $Z = (Z_1, Z_2, \ldots, Z_d)$, and (3) how does this limiting object takes into account the dependence structure within each X_i?

Convergence in distribution of random vectors: here the trick is to recall that we have the following definition of univariate convergence in probability (by Portmanteau): $Y_n \xrightarrow{d} Y$ if

$$\mathbb{E}[g(Y_n)] \to \mathbb{E}[g(Y)],$$

for all bounded continuous g. Note that this "univariate" definition actually immediately generalizes verbatim to random vectors!

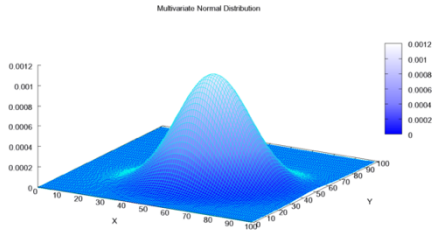
Limiting object: recall that a mean-zero normal can be thought of as a density proportional to exponentiating a polynomial of degree two (with leading coefficient negative to ensure integrability):

$$f(x) \propto \exp(-ax^2) = \exp(-xax),$$

for $a > 0$. The density of the limiting object is a generalization of the above given by

$$f(x) \propto \exp(-x'Ax),$$

where x' is a transpose of dimension 1 by d, A is d by d, and x is d by 1, therefore the product is a scalar. Here is what the density looks like:



To ensure integrability, we ask that the matrix A be such that $x'Ax$ be positive for all $x \in \mathbb{R}^d$, a condition called *positive definiteness*. One fact we will assume from linear algebra is that A being positive definite implies it is invertible.

Multivariate normal: it turns out the reparameterization $\Sigma = 2A^{-1}$, called the covariance matrix, is more interpretable, because then $\mathbb{E}[Z_d Z_{d'}] = \Sigma_{d,d'}$. Moreover, the normalization constant has a closed form, which yields the standard expression for the *multivariate normal distribution*:

$$f(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2} x'\Sigma^{-1}x\right).$$

How to compute Σ in the multivariate CLT: $\mathbb{E}[Z_d Z_{d'}] = \mathbb{E}[X_{1,d} X_{1,d'}]$, i.e. again we can just look at matching the first two moments of one random vector.

Proving tool: Cramér-Wold device. Let Y and Y_n be any random vectors. Then $Y_n \xrightarrow{d} Y$ if and only if for all scalar $v \in \mathbb{R}^d$, $Y_n' v \xrightarrow{d} Y'v$.

6.23 CLT: self-centered version

So far we have assumed the X_i have zero mean. This can always be done without loss of generality but it is often useful to memorize the formula that does that for you:

$$\sqrt{n}(\bar{X}_n - \mu) \to N(0, \Sigma),$$

where the d dimensional random vector $\bar{X}_n$ is the *empirical average*, $\bar{X}_n = (X_1 + X_2 + \cdots + X_n)/n$, the d-dimensional vector μ is the mean, $\mu = \mathbb{E}[X_1]$,

and the matrix Σ is the covariance matrix, which in the non-centered case is $\Sigma_{d,d'} = \mathbb{E}[(X_{1,d} - \mu_d)(X_{1,d'} - \mu_{d'})]$.

6.24 Delta method

Motivation. Recall that the sample variance is given by

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

and has the property that $S_n^2 \xrightarrow{a.s.} \mathbf{Var}X_1$. Can we use the CLT to justify a normal approximation of the distribution of S_n^2 for n large? Convince yourself that it does not "quite" fit the statement of the CLT.

Idea: start with a CLT on the 2-d vectors $Y_i = (X_i, X_i^2)$:

$$\sqrt{n}(\bar{Y}_n - \mu) \to N(0, \Sigma),$$

where, using the shorthand $\alpha_j = \mathbb{E}X_1^j$, $\mu = (\alpha_1, \alpha_2)$, $\Sigma_{1,1} = \alpha_2 - \alpha_1^2$, $\Sigma_{1,2} = \Sigma_{2,1} = \alpha_3 - \alpha_1\alpha_2$, and $\Sigma_{2,2} = \alpha_4 - \alpha_2^2$. Then we will transform this CLT into the convergence statement we want, using a differentiable function ϕ.

Example: $S_n^2 = \phi(\bar{Y}_n)$, where $\phi(x, y) = y - x^2$.

Goal: get a new asymptotic approximation on $\phi(\bar{Y}_n)$ given a known asymptotic approximation on $\bar{Y}_n$. We already know from the continuous mapping theorem that $\phi(\bar{Y}_n) \xrightarrow{a.s.} \phi(\mu)$. This suggests looking at convergence in distribution of

$$\sqrt{n}(\phi(\bar{Y}_n) - \phi(\mu)). \tag{7}$$

How to tackle the weak limit of this sequence of random variables?

Technique: Taylor expansion of ϕ around the mean μ,

$$\phi(\bar{Y}_n) = \phi(\mu) + \phi'(\mu)(\bar{Y}_n - \mu) + \dots.$$

Informally plugging the above into Equation (7), we get

$$\sqrt{n}(\phi(\bar{Y}_n) - \phi(\mu)) \approx \phi'(\mu)\ \sqrt{n}(\bar{Y}_n - \mu) \approx \phi'(\mu)T.$$

Theorem (delta method): if $\phi : \mathbb{R}^k \to \mathbb{R}^m$ is differentiable at μ, and

$$\sqrt{n}(T_n - \mu) \xrightarrow{d} T$$

then

$$\sqrt{n}(\phi(T_n) - \phi(\mu)) \xrightarrow{d} \phi'(\mu)T,$$

where ϕ' is the m by k matrix of partial derivatives, $\phi'_{i,j} = \partial_i\phi_j$.

Proof: if assuming continuous differentiation, a simple application of the mean value theorem, continuous mapping and Slutsky. See e.g. van der Vaart for the full version.

Exercise: complete the motivation, showing that,

$$\sqrt{n}(S_n^2 - \mathbf{Var}X_1) \xrightarrow{d} Z,$$

where Z is a normal with mean zero and variance $\alpha_4 - \alpha_2^2$.

6.25 LLNs and CLTs under relaxed assumptions

Removing all independence assumptions: so far we have allowed dependences between components of random vectors but not across separate random vectors. This can be considerably relaxed. One of the most useful relaxation is based on *Markov chains*. We sketch here the results in their simplest possible form, and will talk in more depth later.

Markov chains, discrete case: consider a list of random variables $X_1, X_2, \ldots$ with a distribution constructed as follows:

$$\mathbb{P}(X_1 = x) = v_x,$$

where $v = (v_1, v_2, \ldots, v_N)$ is a known vector called a initial distribution, $x \in \{1, 2, \ldots, N\}$, and

$$\mathbb{P}(X_i = x' | X_{i-1} = x) = M_{x,x'},$$

where $M = (M_{x,x'})$ is a known N by N matrix called the transition matrix.

Theorem: if M^N has no zero entries,[12] there exists a distribution π such that we have the following LLN:

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \xrightarrow{a.s.} \mu = \int f(x)\pi(\,\mathrm{d}x).$$

and CLT:

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} Z,$$

where $Z \sim N(0, \sigma^2)$.

What is π? This time it is more complicated than the distribution of one of the X_i's, since the distributions are not identical! But it turns out we can still get a lot of knowledge about π, called the *stationary distribution*. Similarly for σ^2, called the asymptotic variance. To be continued when we talk about Markov chains in more depth.

6.26 Convergence in L^p

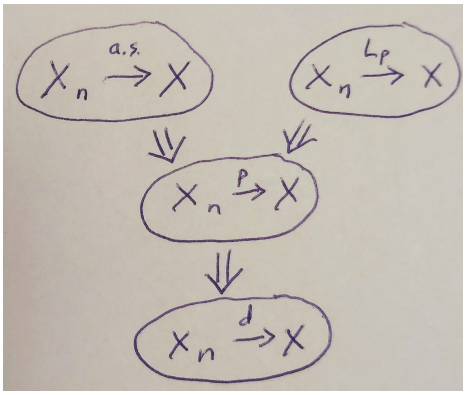
Definition: $\mathbb{E}|X_n - X|^p \to 0$, denoted $X_n \xrightarrow{\mathbf{L}^p} X$. Special cases are called "convergence in mean" (for $p = 1$) or "convergence in mean square" (for $p = 2$).

[12] There are far more general conditions, but they need more setup. We will get to the more general conditions later.

Exercise: show $X_n \xrightarrow{\mathbf{L}^p} X$ implies $X_n \xrightarrow{\mathbb{P}} X$. Hint: use Markov's inequality.

Exercise: the converse is not true. Hint: use the example from Section 3.10.

Hierarchy: This extends Section 6.13 into:



Partial converse: back to the relation between convergence in probability and convergence in mean. For $p = 1$, we have $X_n \xrightarrow{\mathbb{P}} X$ implies $X_n \xrightarrow{\mathbf{L}^p} X$ if and only if they are *uniformly integrable*, defined as: for all $\epsilon > 0$, there is $K > 0$ such that $\sup \mathbb{E}[|X_n|\mathbf{1}(|X_n| \geq K)] \leq \epsilon$.

Exercise: check that indeed the example from Section 3.10 is not uniformly integrable.

7 Poisson theory

7.1 Poisson convergence

Motivation: consider the following historical dataset, containing the number of accidental horse-kick deaths per year in the Prussian army in the 1881–1896 period: 7 deaths in 1881, 1 in 1882, 3, 2, 7, 6, 1, 3, 2, 2, 6, 4, 4, 1, 6, 2. Question: can we approximate the probability that there are no deaths in 1897?

Proposition: if for some constant $\lambda > 0$, $X_n \sim \text{Bin}(n, \lambda/n)$ and $X \sim \text{Poi}(\lambda)$, then $X_n \xrightarrow{d} X$.

Proof: a short proof consists in using characteristic functions. As an exercise, show that the characteristic functions of the binomial random variables in the sequence are given by

$$\varphi_{X_n}(t) = \left(1 - \frac{\lambda}{n} + \frac{\lambda}{n}e^{it}\right)^n,$$

and the characteristic function for the Poisson is given by:

$$\varphi_X(t) = \exp(\lambda(e^{it} - 1)).$$

It is easy to show that for all t, $\varphi_{X_n}(t) \to \varphi_X(t)$.

Exercise: show how to solve the motivation problem using this proposition.

Solution: for a Poisson distribution, the maximum likelihood and moment matching coincide, so we can fit the parameter as follows $\hat{\lambda} \approx \frac{1}{16}(7+1+\cdots+2) = 3.5625$. This yields

$$\mathbb{P}(X = 0) = e^{-\hat{\lambda}}\hat{\lambda}^0/0! \approx 2.8\%.$$

Generalization using "triangular arrays:" let $X_{n,m}$ denote an array with the variables in each row being independent:

- $X_{2,1}, X_{2,2}$ are independent.
- $X_{3,1}, X_{3,2}, X_{3,3}$ are independent.
- etc.

Assume $X_{n,j} \sim \text{Bern}(p_{n,j})$, where

1. we have convergence of row sums to λ: $\lim_{n\to\infty} \sum_{j=1}^{n} p_{n,j} = \lambda$,
2. and as we look at larger and larger row indices n the Bernoulli probabilities become uniformly rare across the columns j: $\lim_{n\to\infty} \max_j p_{n,j} = 0$.

Then: $S_n = X_{n,1} + X_{n,2} + \cdots + X_{n,n}$ is such that $S_n \xrightarrow{d} X$, where $X \sim \text{Poi}(\lambda)$.

Exercise: show that the earlier result is a special case of this one.

Proof: based on the Stein-Chen coupling method. See textbook p. 459.

7.2 Poisson processes: motivation, definition and construction

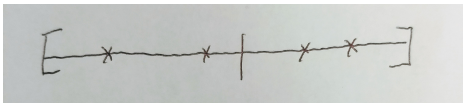
Prerequisite definition: a *Radon-Nikodym derivative* (RN) is like a density, but where we build a measure instead of a probability measure. Compare:

- The probability measure ν has density $f \geq 0$ with respect to a measure μ if
$$\nu(A) = \int_A f(x)\mu(\,\mathrm{d}x).$$
- The measure ν has RN derivative $f \geq 0$ with respect to a measure μ if
$$\nu(A) = \int_A f(x)\mu(\,\mathrm{d}x).$$

Poisson process—motivation and connexion to Poisson convergence:[13]

[13] Reference for the Poisson process: the best reference in my opinion for this part of the material is the beautiful short monograph by Kingman (1993), "Poisson Processes."

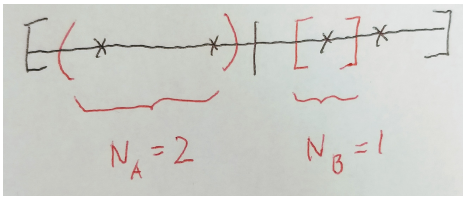
- Let n denote the number of nucleotides in your genome (about 3e9 base pairs).
- Say you are exposed to a small dose of radiation.
- Let Y_i denote the indicator that nucleotide i in a cell mutates.
- Say $Y_i \overset{\text{iid}}{\sim} \text{Bern}(1e-9)$, and let $S_n = Y_1 + Y_2 + \cdots + Y_n$.
- By the Poisson convergence section, S_n is approximately Poisson distributed, with rate parameter λ.
- Now let us view the genome as one line segment $[0, 3]$ (a continuous approximation of the discrete DNA strand).
- What is the approximate probability of the number of mutation in the first half of the genome? Last quarter? Are those independent?



Poisson distribution vs. process: the distribution only keeps track of the total number of mutation, while the process keeps track of where they occur (equivalently, the number in *any* subset of genome).

Notation:

- $\mathcal{X}$: the space in which each point sits in (for example, the interval $[0, 3]$ approximating the genome).
- For $A \subset \mathcal{X}$, let N_A denote the random number of points that fall in A.



From the motivating example: we are interested in cases where:

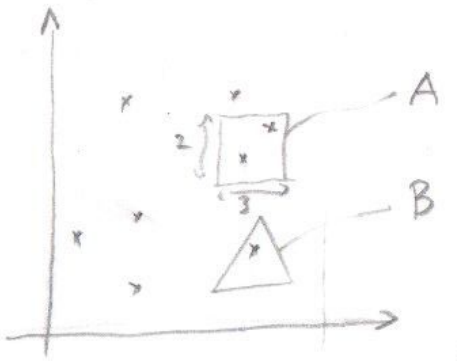
1. $A \cap B = \emptyset \implies N_A$ and N_B are independent.
2. $N_A \sim \text{Poi}(\text{length of } A)$. Let us denote the length of A by $\mu(A)$.

Recall: a nice property of these random variables is:

$$N_{A \sqcup B} = N_A + N_B \sim \text{Poi}(\underbrace{\mu(A) + \mu(B)}_{\mu(A \sqcup B)})$$

Note: this is an instance of what is called *Kolmogorov consistency*.

Example 2: random positions of animals in the forest $\mathcal{X} = [0,1] \times [0,1]$, $\mu = \mu^2$. Exercise: what are the assumptions we implicitly make if we say the animal positions are distributed according to a Poisson process?

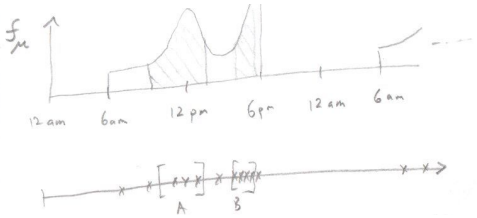


Assumptions in above example:

1. Solitary animals.
2. Do not avoid each other.
3. No food/water source attracting animals.

We will next discuss how to relax the last assumption, in the context of another 1d example.

Example 3: customers entering a store. Here, $\mathcal{X} = [0, \infty)$ represents time. Problem: same expected number of customers between 1am–2am and 1pm–2pm. Solution? Non-uniform μ, defined via a RN derivative.



Compute the intensity via

$$\mu(A) = \int_A f_\mu(x)\,\mathrm{d}x.$$

Then as before,

$$N_A \sim \mathrm{Poi}(\mu(A)).$$

Next: a formal definition of Poisson processes.

Assumptions on μ, called the intensity measure:

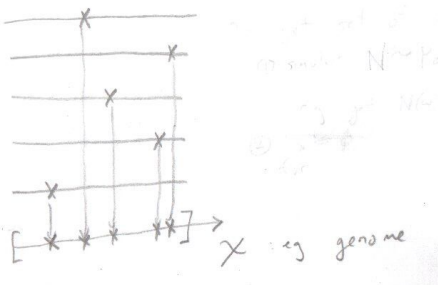
1. A measure on $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$,
2. Can be broken into a countable collection of finite measure segments: exists countable partition A_i of $\mathcal{X}$ such that $\mu(A_i) < \infty$ (this property is called "μ is σ-finite").
3. The intensity measure is non-atomic: for all $x \in \mathcal{X}$, $\mu(\{x\}) = 0$.

Formal definition of a Poisson Process (PP): under the above assumption, the collection of random variables $\{N_A : A \in \mathcal{F}_{\mathcal{X}}\}$ is called a Poisson process with intensity μ if:

1. $A \cap B = \emptyset \implies N_A$ and N_B are independent.
2. $N_A \sim \text{Poi}(\mu(A))$.

Constructive definition:

1. Simulate the point for one of the block A_i at the time:
   (a) Simulate the number of points in A_i, $N_{A_i} \sim \text{Poi}(\mu(A_i))$.
   (b) For $j = 1, \ldots, N_{A_i}$:
      i. $X_{i,j}$ is sampled independently according to μ restricted to A_i and renormalized in order to form a probability distribution, i.e.:
      $$\mathbb{P}(X_{i,j} \in B) = \frac{\mu(B \cap A_i)}{\mu(A_i)}.$$
   (c) The random set of points in A_i is $S_i = \{X_1, \ldots, X_{N_{A_i}}\}$
2. Return the union S of the points over all the S_i's



Proof sketch: the distribution of the algorithm satisfies our two defining axioms.

Lemma: the distribution of S does not depend on the choice of partition $\{A_i\}$.

Proof of lemma: consider two partitions $\{A_i\}$ and $\{B_i\}$. We want to show $S_{\{A_i\}} \stackrel{d}{=} S_{\{B_i\}}$. Build a new partition as we did in exercise 4.7

$$\{C_{i,j} = A_i \cap B_j\}.$$

Show

$$S_{\{A_i\}} \stackrel{d}{=} S_{\{C_{i,j}\}}$$

and

$$S_{\{B_i\}} \stackrel{d}{=} S_{\{C_{i,j}\}},$$

where we $S_{\{A_i\}}$ denote the output of the above algorithm using partition $\{A_i\}$. The key point in the argument is that

$$N_{A_i} \stackrel{d}{=} \sum_j N_{C_{i,j}},$$

which holds by combining additivity of measures

$$\mu(A_i) \stackrel{d}{=} \sum_j \mu(C_{i,j}),$$

and "additivity" of Poisson distributions,

$$X_i \overset{\text{iid}}{\sim} \text{Poi}(\lambda_i) \Longrightarrow \sum_i X_i \sim \text{Poi}\left(\sum_i \lambda_i\right).$$

Back to proposition: given the two set A and B in the axiom, just pick $A_1 = A$ and $A_2 = B$, the proof then follows directly from the lemma.

Counts and sets: two equivalent views. The algorithm shows that the PP can be viewed as a random set of points, denoted $S \sim \text{PP}(\mu)$. With a slight abuse of notation, if μ has RN density f_μ, we might also write $S \sim \text{PP}(f_\mu)$.

7.3 Poisson process with constant intensity on the real line

Special case: linking our definition of PP to the undergraduate definition. Assume:

- $\mathcal{X} = [0, \infty)$
- $\mu =$ uniform.

Let $T_1, T_2, \ldots$ denote the arrival times, i.e. $T_1 = \inf\{t : N_{[0,t]} \geq 1\}$, and more generally, $T_k = \inf\{t : N_{[0,t]} \geq k\}$. Then:

$$\begin{aligned} \mathbb{P}(T_1 > t) &= \mathbb{P}(N_{[0,t]} = 0) && (8) \\ &= \frac{e^{t-0}(t-0)^n}{n!} \quad \text{with } n = 0 && (9) \\ &= 1 - \text{CDF of an exponential} && (10) \end{aligned}$$

Similarly, each inter-arrival times $T_i - T_{i-1}$ are also exponential(1).

7.4 Superposition

Motivation: consider the example concerned with animals in the forest. What if we consider two species simultaneously, assuming they do not interact?

Proposition (superposition): let $\{S_i\}_{i=1}^n$ denote a collection of PPs, $S_i \sim \mathrm{PP}(\mu_i)$, where the intensity measures are defined on a common space, $\mu_i : \mathcal{F}_\mathcal{X} \to [0, \infty)$. Then $S \coloneqq \cup_i S_i \sim \mathrm{PP}(\sum_i \mu_i)$.

Exercise: prove this using the algorithmic construction and the fact that the sum of Poissons is Poisson.

Note: this can be generalized to countable unions of PPs, provided that the sum of intensities is still σ-finite.

7.5 Thinning

Definition (thinning). Let $S = \{X_1, X_2, \dots\} \sim \mathrm{PP}(\mu)$, $Y_i \overset{\text{iid}}{\sim} \mathrm{Bern}(p)$. We define the thinned process by using the iid coin flips Y_i to decide, for each point in S, whether we keep this point or not: $T \coloneqq \{X_i \in S : Y_i = 1\}$.

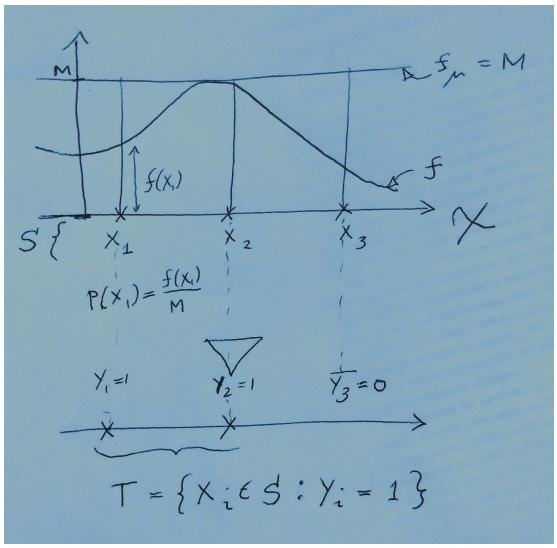
Proposition (thinning). The random set T defined above by the thinning procedure is a Poisson process, and its intensity is obtained by scaling down the original intensity μ: $T \sim \mathrm{PP}(p\mu)$, where $(p\mu)(A) \coloneqq p\mu(A)$.

Proof: will be easier once we talk about conditioning next week.

Thinning for densities. Suppose now the intensity measure has a RN derivative, $S = \{X_1, X_2, \dots\} \sim \mathrm{PP}(f_\mu)$. We can let the coin flips be non-identical: given a position of a point x, let us say we use for the corresponding Y a coin with success probability $p(x)$ for some $p : \mathcal{X} \to [0, 1]$. Then $T \coloneqq \{X_i \in S : Y_i = 1\}$ is again a PP, this time with intensity RN given by the pointwise product $f(x) = p(x) f_\mu(x)$.

Powerful computational trick: let us say you have a complicated intensity function with RN f from which you would like to simulate a PP. If f is bounded by M, then you can use the following recipe:

1. Let $f_\mu = M$, the uniform density scaled by M. Note: $f(x) = p(x) f_\mu(x)$ for some function $p : \mathcal{X} \to [0, 1]$, namely
$$p(x) = \frac{f(x)}{f_\mu(x)} = \frac{f(x)}{M}.$$
2. Simulate $S = \{X_1, X_2, \dots\} \sim \mathrm{PP}(f_\mu)$ which is trivial to do.
3. For each X_i, simulate a Bernoulli Y_i with success probability $p(X_i)$.
4. Return $T \coloneqq \{X_i \in S : Y_i = 1\}$. By the second thinning theorem, this is distributed according to a PP with intensity f as hoped.

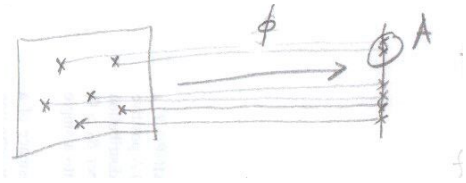


Exercise: generalize this recipe to $f_\mu(x)$ being some non-constant upper-bound on f.

7.6 Mapping

Motivation: consider the animals in the forest example. What is the distribution of the x-coordinates of the random animal locations?

More general setup: suppose we have a mapping $\phi : \mathcal{X} \to \mathcal{X}'$.



Exercise: write the projection example using the framework of a mapping ϕ.

Solution: In the above example, $\mathcal{X} = [0,1]^2$, $\mathcal{X}' = [0,1]$, and $\phi(x,y) = x$.

Proposition: Let $S \sim \mathrm{PP}(\mu)$. Assume $\mu^\star(A) \coloneqq \mu(\phi^{-1}(A))$ has no atom and is σ-finite. Then $\phi(S) \sim \mathrm{PP}(\mu^\star)$.

7.7 Compound PP

Observation: let us view the collection of random variable N_A differently, writing it as $N(\cdot)$ instead. What is this? A random measure! Now let us fix a

real-valued function $f : \mathcal{X} \to \mathbb{R}$. We can then define the random integral:

$$Y = \int f \, \mathrm{d}N = \sum_{X \in S} f(X).$$

Definition: Y is called a compound Poisson process.

Next: we would like to gain information about the distribution of Y. We will see that we can compute its characteristic function:

$$\mathbb{E}\left[e^{itY}\right] = \exp\left\{\int \left(e^{itf(x)} - 1\right) \mu(\,\mathrm{d}x)\right\}.$$

Proof: start with f simple:

$$f = \sum_j a_j \delta_{A_j}.$$

Let $\lambda_j = \mu(A_j)$, $N_j \sim \mathrm{Poi}(\lambda_j)$, and

$$Y = \sum_j a_j N_j.$$

We now compute the characteristic function as follows:

$$\begin{aligned}
\mathbb{E}[e^{itY}] &= \prod_j \mathbb{E}[e^{ita_j N_j}] \\
&= \prod_j e^{\lambda_j (e^{ita_j} - 1)} \\
&= \exp\left(\sum_j \lambda_j (e^{ita_j} - 1)\right) \\
&= \exp\left(\int (e^{itf(x)} - 1)\mu(\,\mathrm{d}x)\right).
\end{aligned}$$

Next, for arbitrary f, use DCT to obtain:

$$\mathbb{E}[e^{itY}] = \exp\left(\int (e^{itf(x)} - 1)\mu(\,\mathrm{d}x)\right).$$

8 Conditioning

8.1 Background: σ-algebra and information

It will be useful for you to review Section 2.27 before reading this chapter.

8.2 Conditioning on an event

Motivation: a couple has two children. You would like to predict the sex of the second (youngest) child. Your initial beliefs over the sex of that second child is 1/2 boy, 1/2 girl. You get one piece of information: at least one of the two children is a girl. What is the optimal way of updating your beliefs?

Conditioning on an even: consists in an operator that takes as an input:

1. some *a priori* beliefs (a probability distribution $\mathbb{P}(\cdot)$),
2. as well as an observed event E.

The interpretation of E is that you know the true outcome is somewhere in E, but you still do not know which outcome in E. The output of conditioning on an event: a new, updated belief, denoted $\mathbb{P}(\cdot|E)$. The optimal value for this updated belief, for any query set A, is given by:

$$\mathbb{P}(A|E) = \frac{\mathbb{P}(A \cap E)}{\mathbb{P}(E)}.$$

Example. In the motivation at the beginning of the section, we are interested in A and $A^{\complement}$, where $A = \{(g,g),(b,g)\}$. The observation is $E = \{(b,g),(g,b),(g,g)\}$, hence the updated belief for the motivation question is $\mathbb{P}(A|E) = 2/3, \mathbb{P}(A^{\complement}|E) = 1/3$.

8.3 Conditioning on a random variable

Interpretation: let X and Y be two random variables defined on the same probability space:

- Y is observed, (e.g. $Y = \mathbf{1}_E$ in the motivating example)
- X is the variable you would like to predict (e.g. $X = \mathbf{1}_A$ in the motivating example).

The conditional expectation of X given Y is a new estimator random variable $\delta = f(Y)$, which can be interpreted as the "best" estimator of X based on Y. Notation: this random variable δ is denote by $\mathbb{E}[X|Y]$.

Notion of optimality: consider the optimization program

$$\text{minimize } \mathbb{E}(f(Y) - X)^2,$$

where minimization is performed over functions f (technically, over measurable functions). If a function f^* maximizes this program and give a finite value to the objective function, then $\mathbb{E}[X|Y] \coloneqq f^*(Y)$.

Exercise: show that in the two children problem, you obtain:

$$f^*(y) = \begin{cases} 0 & \text{if } y = 0 \\ 2/3 & \text{if } y = 1 \\ \text{any value} & \text{otherwise.} \end{cases}$$

Solution: we have:

$$\begin{aligned}\mathbb{E}(f(Y)-X)^2 &= \frac{1}{4}\left[(f(0)-0)^2+(f(1)-1)^2+(f(1)-0)^2+(f(1)-1)^2\right]\\ &= \frac{f(0)^2+3f(1)^2-4f(1)+2}{4},\end{aligned}$$

hence we see right away that $f(0)=0$ and that the values of f are unimportant except at $y=0$ and $y=1$. To find $f_1=f(1)$, we compute the derivative with respect to f_1 and get

$$\frac{\mathrm{d}}{\mathrm{d}f_1}\left(\frac{3f_1^2-4f_1+2}{4}\right)=\frac{3}{2}f_1-1,$$

hence finding the root we get indeed $f_1=2/3$.

Note: by the discussion in Section 8.1, note that we only used Y via $\sigma(Y)$. For this reason, we can define the notion of conditioning on a σ-algebra $\mathcal{F}$, denoted $\mathbb{E}[X|\mathcal{F}]$, in the same way as we did above.

Notation: $\|X\|_2 \coloneqq \sqrt{\mathbb{E}[X^2]}$, $\mathbf{L}_2 \coloneqq \{\text{r.v. } X : \|X\|_2 < \infty\}$.

Note: we will see that the above definition always works when $Y \in \mathbf{L}_2$. To generalize this to $Y \in \mathbf{L}_1$, we will make use of a generalization of a property which is itself very important in practice: the law of total expectation.

8.4 The law of total probability and expectation

Useful property: the law of total expectation:

$$\mathbb{E}[X]=\mathbb{E}[\mathbb{E}[X|Y]],$$

and its associated law of total probability, a special case where $X=\mathbf{1}_A$:

$$\mathbb{P}(X\in A)=\mathbb{E}[\mathbb{P}(X\in A|Y)].$$

Example/exercise: suppose that

$$\begin{aligned}U &\sim \mathrm{Unif}(0,1)\\ X|U &\sim \mathrm{Bin}(n,U),\end{aligned}$$

Show that $\mathbb{P}(X=i)=1/(n+1)$ using the law of total probability. Hint: use the Beta function, defined for $x>0$ and $y>0$:

$$B(x,y)=\int_0^1 u^{x-1}(1-u)^{y-1}\,\mathrm{d}u=\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)},$$

where Γ denotes the Gamma function, which for integer x satisfies $\Gamma(x)=(x-1)!$.

Notation used above: what do we need precisely by "$X|U \sim \text{Bin}(n, U)$"? More generally, suppose $\{D_y\}$ is some collection of distribution indexed by a parameter y (e.g. the Binomials indexed by their success probability parameter). Then the notation "$X|Y \sim D(Y)$" means that for all A, $\mathbb{P}(X \in A|Y) = D_Y(A)$.

Solution: by the Law of Total Probability, $\mathbb{P}(X = i) = \mathbb{E}[\mathbb{P}(X = i|U)]$. By the distributional statement $X|U \sim \text{Bin}(n, U)$, and the fact that the Binomial distribution has a known probability mass function in which we plug in (compose) with the random success probability parameter U:

$$\begin{aligned}\mathbb{P}(X = i) &= \mathbb{E}[\mathbb{P}(X = i|U)] \\ &= \mathbb{E}\left[\binom{n}{i} U^i (1-U)^{n-i}\right] \\ &= \binom{n}{i} \mathbb{E}\left[U^i (1-U)^{n-i}\right].\end{aligned}$$

Now the last line is the expectation of a function $g(u) = u^i(1-u)^{n-i}$ of a uniform random variable. Therefore we can compute as:

$$\begin{aligned}\mathbb{P}(X = i) &= \binom{n}{i} \mathbb{E}\left[U^i (1-U)^{n-i}\right] \\ &= \binom{n}{i} B(i+1, n-i+1) \\ &= \frac{1}{n+1},\end{aligned}$$

using the hint.

Discrete case: to get more intuition on the law of total expectation/probability, let us first assume that Y is simple, $Y = \sum_{i=1}^n a_i \mathbf{1}_{A_i}$, where A_i forms a partition. By simple set properties:

$$\begin{aligned}\mathbb{P}(A) &= \sum_{i=1}^n \mathbb{P}(A \cap (Y = a_i)) \\ &= \sum_{i=1}^n \mathbb{P}(A|Y = a_i)\mathbb{P}(Y = a_i).\end{aligned}$$

Now, let:

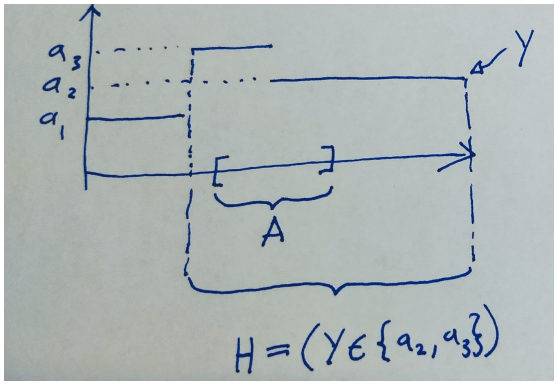
$$f(y) = \begin{cases} \mathbb{P}(A|Y = y) & \text{if } y \in \{a_1, \ldots, a_n\} \\ \text{arbitrary,} & \text{otherwise.} \end{cases} \tag{11}$$

Conditioning on a random variable can be thought as introducing a nice notation for this identity, $\mathbb{P}(A|Y) := f(Y)$. This also provides a way to generalize to conditioning into a non-discrete random variable.

Next: we first generalize the above to cases where we select only a subset of the blocks in the partition. To formalize this, let $B \subset \{a_1, \ldots, a_n\}$, then a set

of blocks can be denoted as $H \coloneqq (Y \in B)$, i.e. $H \in \sigma(Y)$. With this notation, show that:

$$\mathbb{P}(A \cap H) = \mathbb{E}[\mathbf{1}_H \mathbb{P}(A|Y)].$$



Definition of conditional probability: we will use this property, which we have shown to hold for simple Y, as the basis of our fully general definition of conditional probability. Let $Y \in \mathbf{L}_1$. Then there exists a random variable, denoted $\mathbb{P}(A|Y)$, such that:

1. $\mathbb{P}(A|Y) \in \sigma(Y)$, (i.e. $\mathbb{P}(A|Y)$ is an estimator based only on the information offered by Y)
2. for all $H \in \sigma(Y)$, $\mathbb{P}(A \cap H) = \mathbb{E}[\mathbf{1}_H \mathbb{P}(A|Y)]$.

Moreover this random variable is almost sure unique.

Ingredient for the proof: Radon-Lebesgue-Nikodym theorem let $(\Omega, \mathcal{G})$ denote a measurable space with two measures ν and μ. Then $\mu(A) = 0 \Rightarrow \nu(A) = 0$ (a property called "ν is absolutely continuous with respect to μ", denoted $\nu \ll \mu$), if and only if ν has a density with respect to μ, i.e. $\nu(A) = \int_A f \, \mathrm{d}\nu$ for some $f \in \mathcal{G}$.

Proof of existence of conditional expectation: let $\mathcal{G} = \sigma(Y)$, and define, for all $H \in \mathcal{G}$, $\nu(H) = \mathbb{P}(A \cap H)$ and $\mu(H) = \mathbb{P}(H)$. Since $\mu(H) = 0 \Longrightarrow \nu(H) = 0$, it follows by Radon-Lebesgue-Nikodym that there is a $f \in \mathcal{G}$ such that $\nu(H) = \int_H f \, \mathrm{d}\mu$. This yields the result with $f = \mathbb{P}(A|Y)$.

Proof of uniqueness: suppose there are two random variables $Z_1 = \mathbb{P}(A|Y)$ and another one Z_2 both satisfying the two conditions in the definition of conditional probability. We have $\mathbb{E}[\mathbf{1}_H Z_1] = \mathbb{P}(A \cap H) = \mathbb{E}[\mathbf{1}_H Z_2]$, therefore $\mathbb{E}[\mathbf{1}_H (Z_1 - Z_2)] = 0$ for all $H \in \sigma(Y)$. Take $H = (D > 0)$, we get by Chebyshev, $\mathbb{E}[\mathbf{1}_H D] = 0 \Longrightarrow \mathbb{P}(D > 0) = 0$. By a symmetrical argument on $Z_2 - Z_1$, $\mathbb{P}(D < 0) = 0$.

Definition of conditional expectation: a generalization of the above yield also the definition conditional expectation. Let $X, Y \in \mathbf{L}_1$ be defined on a common probability space. Then there exists a random variable, denoted $\mathbb{E}[X|Y]$, such that:

1. $\mathbb{E}[X|Y] \in \sigma(Y)$, (i.e. $\mathbb{E}[X|Y]$ is an estimator of X based on Y)
2. for all $H \in \sigma(Y)$, $\mathbb{E}[\mathbf{1}_H X] = \mathbb{E}[\mathbf{1}_H \mathbb{E}[X|Y]]$.

Moreover this random variable is almost sure unique.

8.5 Key properties

Important exercise: suppose (X, Y) have joint density $f(x, y)$. Then $\mathbb{E}[X|Y] = \psi(Y)$, where

$$\begin{aligned} \psi(y) &:= \begin{cases} \int x f_{X|Y}(x|y)\,\mathrm{d}x & \text{if } f_Y(y) > 0 \\ \text{arbitrarily,} & \text{otherwise,} \end{cases} \\ f_{X|Y}(\cdot|y) &\propto f(\cdot, y) \\ &:= \frac{f(\cdot, y)}{f_Y(y)} \\ f_Y(y) &:= \int f(x, y)\,\mathrm{d}x. \end{aligned}$$

Terminology: the function $f_Y(\cdot)$ is called the marginal density, and the function $f_{X|Y}(\cdot|y)$, the conditional density. Note that these are indeed densities.

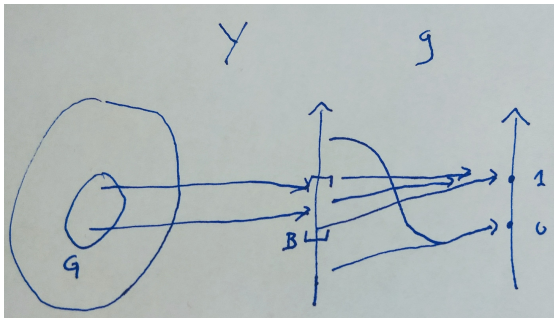
Property: when $X, g(Y) \in \mathbf{L}_1$,

$$\mathbb{E}[Xg(Y)|Y] = g(Y)\mathbb{E}[X|Y].$$

Proof: we have to show that the RHS satisfies the two axioms of conditional expectation. For (1), clearly it is of the form $f(Y)$. For (2), let $H \in \sigma(Y)$, and suppose first that $g(y) = \mathbf{1}_B$, implying that $g(Y) = \mathbf{1}_G$, $G \in \sigma(Y)$. We get:

$$\begin{aligned} \mathbb{E}[\mathbf{1}_H X \mathbf{1}_G] &= \mathbb{E}[\mathbf{1}_H \mathbf{1}_G \mathbb{E}[X|Y]] \Leftrightarrow \\ \mathbb{E}[\mathbf{1}_{H'} X] &= \mathbb{E}[\mathbf{1}_{H'} \mathbb{E}[X|Y]], \end{aligned}$$

which is true since $H' \in \sigma(Y)$. Complete the proof using the DCT.



8.6 Equivalence of the two definitions

So far, we have given two definitions of conditional expectations, one that works for $\mathbf{L}_2$ only, and one that works for $\mathbf{L}_1$ and $\mathbf{L}_2$. Here we show that for the $\mathbf{L}_2$ setup, our new, general definition agrees with the $\mathbf{L}_2$ specific earlier definition:

$$\begin{aligned}
\mathbb{E}[(f(Y)-X)^2] &= \mathbb{E}[\mathbb{E}[(f(Y)-X)^2|Y]] \\
&= \mathbb{E}[\underbrace{\mathbb{E}[f(Y)^2|Y]}_{(f(Y))^2} - \underbrace{2\mathbb{E}[f(Y)X|Y]}_{2f(Y)\mathbb{E}[X|Y]} + \mathbb{E}[X^2|Y]] \\
&= \mathbb{E}[\underbrace{(f(Y)-\mathbb{E}[X|Y])^2}_{\geq 0} + \underbrace{\mathbb{E}[X^2|Y]-(\mathbb{E}[X|Y])^2}_{:=\mathbf{Var}[X|Y]}].
\end{aligned}$$

Note:

1. the RHS$\geq \mathbb{E}[\mathbf{Var}[X|Y]]$.
2. if $f(Y)=\mathbb{E}[X|Y]$ a.s., then RHS$= \mathbb{E}[\mathbf{Var}[X|Y]]$.

8.7 The Bayes estimator (a special case)

Motivation: a complete order over estimator.

Frequentist notion of optimality: since in the MLE (Maximum Likelihood Estimation) framework we do not place a prior on $\theta = X$, the performance of an estimator δ depends on the true parameter θ:

$$\mathrm{MSE}_\theta(\delta) = \mathbb{E}_\theta[(\delta-\theta)^2] = \mathbb{E}[(\delta-\theta)^2|\theta].$$

The MLE mimizes this objective function.

Issue: the space of functions (of θ in this case) is not a complete order. There is not a notion of an a.s. best estimator under the above criterion.

Solution: restrict the class of estimators, e.g. to unbiased ones.

Criticism: can be restrictive/artificially rule out good estimators.

Bayesian alternative: average over θ, to get a real number summary of each estimator:

$$\mathrm{mse}(\delta) = \mathbb{E}[\mathrm{MSE}_\theta(\delta)].$$

Consequence: this yields a *complete order* over estimator. The name of the best estimator is called the *Bayes estimator*, and corresponds to the posterior mean $\mathbb{E}[\theta|\text{data}]$ with our choice of loss function being the square difference.

Note: the Bayes estimator can be defined for other loss functions, in which case it involves an optimization problem where the objective function is a conditional expectation given the data.

8.8 Geometric view of expectation and further properties

Suppose in the remaining of the section that $Y \in \mathbf{L}_2$.

Recall:

- Real vector space: a set V of points in $\mathbb{R}^d$ and a $+$ and $\cdot$ operations such that a set of useful properties hold (associativity, $v_1 + (v_2 + v_3) = (v_1 + v_2) + v_3$, distributivity $a(v_1 + v_2) = av_1 + av_2$, etc (see wikipedia)).
- Abstract vector space: a set of objects V with two operations satisfying the same properties.

Example: $V = \mathbf{L}_2 = \{\text{r.v. } X : \Omega \to \mathbb{R}, \mathbb{E}X^2 < \infty\}$. Easy exercise: read the axioms on wikipedia and check they are satisfied with $+$ denoting the addition of functions and $\cdot$, the multiplication of a function by a constant.

Three important ideas from linear algebra.

1. Norm: $\|\cdot\| : V \to [0, \infty)$. Key defining property (see wiki for the other ones): $\|v + w\| \leq \|v\| + \|w\|$ (triangle inequality).

Examples:

1. $V = \mathbb{R}^2$, $\|v\|_2 = \sqrt{v_1^2 + v_2^2}$, triangle inequality is the Pythagorean theorem.
2. $V = \mathbf{L}_2$, $\|X\| = \sqrt{\mathbb{E}X^2}$, triangle inequality is called Minkowski's inequality, described in more detail shortly.

2. Subspace: a "closed" subset of a vector space, $W \subset V$, i.e. such that $v_1, v_2 \in W \Rightarrow v_1 + v_2 \in W$.

Examples:

1. $\{\text{points of the form } (0, x)\} \subset \mathbb{R}^2$,
2. $W = \{Z \in \mathbf{L}_2 : Z \in \sigma(Y)\} = \{Z \in \mathbf{L}_2 : Z = f(Y), f \text{ measurable}\}$.

3. Projection: of a point v into a subspace W. This projection is defined as:

$$\text{argmin}_{w \in W} \|v - w\|.$$

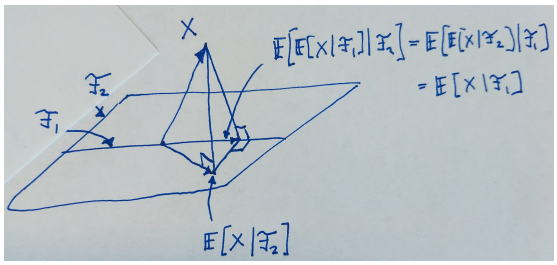
Key example: when $V = \mathbf{L}_2$, W as point 2 above:

$$\begin{aligned} \mathbb{E}[X|Y] &= \text{argmin}_{Z \in W} \|Z - X\| \\ &= \text{projection of } X \text{ into } W. \end{aligned}$$

Application: if $\mathcal{F}_1 \subset \mathcal{F}_2$, then:

$$\begin{aligned} \mathbb{E}[\mathbb{E}[X|\mathcal{F}_1]|\mathcal{F}_2] &= \mathbb{E}[\mathbb{E}[X|\mathcal{F}_2]|\mathcal{F}_1] \\ &= \mathbb{E}[X|\mathcal{F}_1]. \end{aligned}$$

Proof sketch:



Last few properties of conditional expectations:

1. Linearity: $\mathbb{E}[aX + Y|\mathcal{F}] = a\mathbb{E}[X|\mathcal{F}] + \mathbb{E}[Y|\mathcal{F}]$.
2. Monotonicity.
3. Jensen's inequality.
4. Chebyshev.

8.9 Geometric view: more details on triangle inequality

Minkowski inequality: is usually proven using *Cauchy-Schwarz inequality*, which is a must-know itself.

Cauchy-Schwarz: $|\langle u, v\rangle| \leq \|u\|_2\|v\|_2$, where $\langle\cdot,\cdot\rangle$ denotes an inner product, e.g. $\langle X, Y\rangle = \mathbb{E}[XY]$. This result should be pretty intuitive: suppose you have two vectors u and v of fixed length and you want to maximize their dot product. How to do this? Make them point in the same direction! In this case we actually get equality. In the case of a real vector space, this is clear from the identity $\langle u, v\rangle^2 = (\|u\|_2\|v\|_2\cos\theta)^2 \leq (\|u\|_2\|v\|_2)^2$, where θ is the angle between the two vectors. See wikipedia for the general proof.

Minkowski from Cauchy-Schwarz:

$$\begin{aligned}
\|X+Y\|_2^2 &= \mathbb{E}|X+Y|^2 \\
&\leq \mathbb{E}[(|X|+|Y|)|X+Y|] \\
&= \mathbb{E}[|X|\times|X+Y|] + \mathbb{E}[|Y|\times|X+Y|] \\
&\leq \|X\|_2\|X+Y\|_2 + \|Y\|_2\|X+Y\|_2 \quad \text{(using Cauchy-Schwarz)} \\
&= (\|X\|_2 + \|Y\|_2)\|X+Y\|_2.
\end{aligned}$$

Finally, divide each side by $\|X+Y\|_2$.

Generalizations: to go from L^2 to L^p, $p \geq 1$ is easy. Search *Hölder's inequality* and use that instead of Cauchy-Schwarz in the above argument.

8.10 Conditional independence

Conditional independence of events: events A and B are conditionally independent given C, denoted $A \perp\!\!\!\perp B|C$, if

$$\mathbb{P}(A \cap B|C) = P(A|C)P(B|C).$$

Notice this is just the standard notion of independence applied to a conditional probability $\mathbb{P}(\cdot|C)$.

Conditional independence of random variables: two random variables X and Y are conditionally independent given Z, $X \perp\!\!\!\perp Y|Z$, if

$$\mathbb{E}[g_1(X)g_2(Y)|Z] = \mathbb{E}[g_1(X)|Z]\mathbb{E}[g_2(Y)|Z]$$

for all bounded continuous g_1, g_2.

Exercise: independence does not imply conditional independence, and conditional independence does not imply independence. Hint: consider the following

1. The random variables X_1, X_2, X_3 corresponding to the position of a player of ladders and snakes at turns one, two and three (ignore the snakes and ladders for simplicity), $X_1 \sim \text{Unif}(1, \ldots, 6)$, $X_i|X_{i-1} \sim \text{Unif}(X_{i-1} + 1, \ldots, X_{i-1} + 6)$.
2. The random variable $S = Y_1 + Y_2$ where Y_1 and Y_2 correspond to the throwing of two dice.

8.11 Directed graphical models

Generative model: a description of a joint distribution as a product of conditional distribution. By the chain rule, this can always be done, but typically we expect a generative model to be such that each conditional distribution can be efficiently simulated from.

Example: (made up)

$$\begin{aligned}
X_1 &\sim \text{Unif}(0, 1/2, 1) \\
X_2|X_1 &\sim \text{Bern}(X_1) \\
X_3|X_1, X_4 &\sim \text{Poi}(1 + X_1 + X_4) \\
X_4 &\sim \text{Bern}(1/3) \\
X_5|X_3 &\sim \text{Unif}(0, .., X_3) \\
X_7 &\sim \text{Geo}(1/4) \\
X_6|X_4, X_7 &\sim \text{Unif}(X_4, X_7) \\
X_8 &\sim \text{Norm}(0, 1)
\end{aligned}$$

Motivation for directed graphical models: they help us address the following two problems:

1. Design an algorithm to sample from the joint distribution $f(x_1, x_2, \ldots, x_8)$.
2. Efficiently establish some conditional independence relationships. For example, do we have $X_4 \perp\!\!\!\perp X_2 | X_1, X_5, X_6$? We will cover this in the next section.

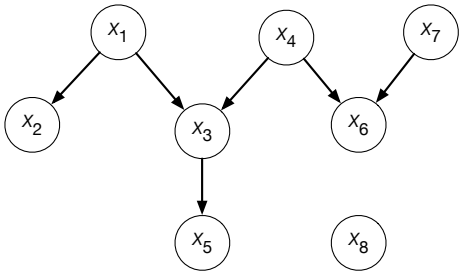
First problem: sampling from the joint distribution. This task arises when you need to generate "synthetic data" in a Bayesian context. This is also called "forward sampling." The basic problem is that if there are n variables, the chain rule can be written in $n!$ ways. Which one to pick? E.g.: compare these three orders

1. $f(x_1, x_2, \ldots, x_8) = f(x_1)f(x_2|x_1)f(x_3|x_1, x_2)f(x_4|x_1, x_2, x_3) \ldots f(x_8|x_1, \ldots, x_7)$, and order we denote by $(1, 2, 3, 4, 5, 6, 7, 8)$,
2. order $(1, 4, 7, 2, 3, 6, 5, 8)$,
3. order $(8, 1, 2, 4, 3, 5, 7, 6)$.

It is much easier to simulate from order 2 than order 1. Why? At the same time, order 2 is not unique: 3 would be just as efficient. Directed graphical models will bring much clarity here:

Definition: a directed graphical model is a directed model where nodes are variables and there is an edge from variable X_i to X_j if the conditional distribution of X_j in the generative model depends on X_i.

Example. For the previous generative model, we obtain:



Forward sampling using graphical models: pick an order $(i_1, i_2, \ldots)$ such that the edges of the graphical model are respected, i.e. for all edge $(i_k \to i_l)$, the first end point appear earlier in the order compared to the second, $k < l$. This is called a *linearization* of a partial order, and can be performed in linear time in the size of the graph, via an algorithm called *topological sorting*.

8.12 Establishing (conditional) independence relations using directed graphical models

Establishing certain independence relations. Let us start with independence, and do conditional independence after.

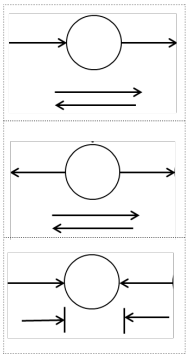
Idea: We can sometimes infer independence relationships just by looking at the shape of the graphical model.

Example: using the same graphical model as before, suppose we want to find if X_4 is independent of X_7.

Bayes ball algorithm for independence. The rules of this algorithm are as follows:

1. Two nodes *communicate* if there is some path following the graphical rules below. In this case we *cannot say* just from the graphical model if the random variables are independent

2. If there are no such path, the nodes do not communicate. In this case we can conclude the random variables are independent



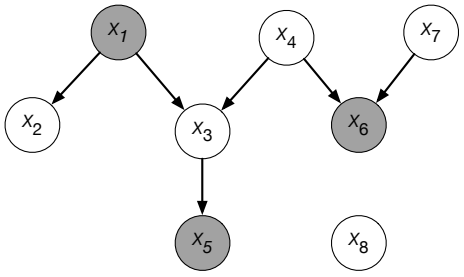
Exercise: In our usual example, which nodes are guaranteed to be independent of X_4?

Solution: X_1, X_2, X_7, X_8.

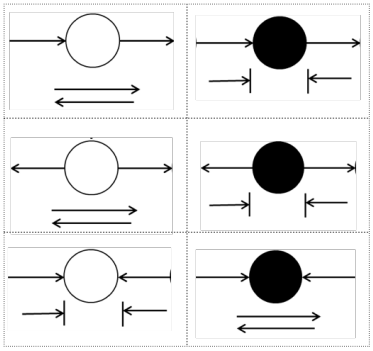
Conditional independence. Suppose now we want to find conditional independence statements. We first extend our notion of graphical model:

Convention: shade in grey the random variables that we want to condition upon.

Example. In our recurring model, let us condition on X_1, X_5, X_6.



Bayes ball algorithm for independence. The algorithm works in the same way but with the following extended rule set taking into account shaded nodes:



Example: From the usual graphical model, which of the white nodes[14] are guaranteed to be conditionally independent of X_4 given the shaded nodes?

Solution: X_2 and X_8.

9 Markov chains

9.1 Basic definitions and examples

Informal: A sequence of random variables where the future is independent of the past given the present.

From the graphical model point of view: a chain-shaped graphical model.

Formal: a sequence of random variables $X_i : \Omega \to \mathcal{X}$ is Markov if for all $A \in \mathcal{F}_{\mathcal{X}}$,

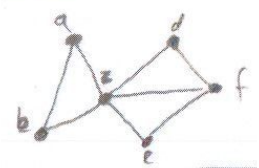
$$P(X_{n+1} \in A | X_0, X_1, \ldots, X_n) = \mathbb{P}(X_{n+1} \in A | X_n).$$

Examples:

1. Let E denote a set of undirected edges over $\mathcal{X}$. Define

$$\mathbb{P}(X_{n+1} = x | X_n) = \frac{1}{Z(X_n)} \mathbf{1}[\{x, X_n\} \in E].$$

Note that the normalization $Z(x)$ is given by the number of nodes connected to x. For example $Z(\mathrm{f}) = 3$ below:



[14] Shaded nodes are always conditionally independent, since, for any events A, B, C: $\mathbb{P}(AB|BC) = \mathbb{P}(A|BC)\mathbb{P}(B|BC) = \mathbb{P}(A|BC)$.

To make this more interesting, consider the vertices of the graph given by the location of a knight on a game of chess (8x8 square), and the edges given by the moves allowed for the knight (L shaped moves moving by one square in one axis, and 2 in the other axis).

2. The Wright-Fisher model: suppose that N bacteria can live in a Petri dish. There are two species, blue and green bacteria. In the first day, there are X_1 green and $N - X_1$ blue bacteria. In the next day, there are still N bacteria, the descendants from the previous generation. The parent of each of the N in day 2 are selected independently and uniformaly from those in day 1. The color is inherited without mutation. This mean:

$$\mathbb{P}(X_n = k|X_n) = \binom{N}{k} \left(\frac{X_n}{N}\right)^k \left(1 - \frac{X_n}{N}\right)^{N-k}.$$

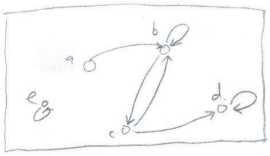
Note that $X_n = 0 \Rightarrow X_{n+1} = 0$, and $X_n = N \Rightarrow X_{n+1} = N$. These are called absorbing states.

9.2 Representation under the homogeneity condition

Note: from the previous part on conditioning, we have that $\mathbb{P}(X_{n+1} = y|X_n) = f(X_n)$ for some $f(x)$. In fact, this function will also depend on y and n. We denote it by $p_n(x \to y)$ and call it the transition probability.

Definition: a Markov chain is homogeneous is $p_n(x \to y) = p_{n+1}(x \to y)$ for all n, x, y. We denote it by $p(x \to y)$.

Visualization: of $p(x \to y)$ via a *state diagram*. Consider a directed graph where the nodes are the points in $\mathcal{X}$ and where the is an edge $(x \to y) \in E$ if and only if $p(x \to y) > 0$. Informally, this encodes the sparsity patterns of the transition probabilities.



Note: if X_n is finite and homogeneous, it is characterized by 2 objects:

1. the transition probabilities, $p = p_n$,
2. an initial distribution with pmf μ, $\mathbb{P}(X = x) = \mu(x)$.

Notation: it is often useful to have the initial distribution put all the mass on a single point x, in which case we write $\mathbb{P}_x$. E.g. $\mathbb{P}_x(X_1 = y) = p(x \to y)$.

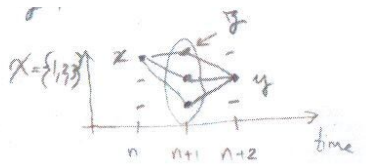
9.3 First connection with linear algebra: Chapman-Kolmogorov equation

Matrix notation for transition probabilities: to create connections with linear algebra, it will be useful to organize the transition probabilities into a matrix (assuming without loss of generality that $\mathcal{X} = \{1, 2, \ldots, K\}$): $M_{x,y} = p(x \to y)$.

Vector notation for pmfs over the states: if μ is a pmf over $\mathcal{X}$ (for example, an initial distribution), we can view it as a vector, $\mu = (\mu(1), \mu(2), \ldots, \mu(K))$.

How to find the 2-step transition, i.e. $\mathbb{P}(X_{n+2} = y | X_n = x)$? As usual we reintroduce the random variable X_{n+1} using marginalization/the law of total probability:

$$\begin{aligned}
\mathbb{P}(X_{n+2} = y | X_n = x) &= \sum_{z \in \mathcal{X}} \mathbb{P}(X_{n+2} = y, X_{n+1} = z | X_n = x) \\
&= \sum_{z \in \mathcal{X}} \mathbb{P}(X_{n+1} = z | X_n = x) \mathbb{P}(X_{n+2} = y | X_n = x, X_{n+1} = z) \quad \text{(chain rule)} \\
&= \sum_{z \in \mathcal{X}} \mathbb{P}(X_{n+1} = z | X_n = x) \mathbb{P}(X_{n+2} = y | X_{n+1} = z) \quad \text{(Markov assumption)} \\
&= \sum_{z \in \mathcal{X}} p(x \to z) p(z \to y) \\
&= (M^2)_{x,y}.
\end{aligned}$$



Exercise: show that more generally,

$$\mu M^n = (\mathbb{P}(X_n = 1), \mathbb{P}(X_n = 2), \ldots, \mathbb{P}(X_n = K)).$$

9.4 Hitting probabilities

Setup: suppose we have an homogeneous finite state Markov chain X_n with exactly two absorbing states, denoted x_1 and x_2.

Example/exercise: create a state space to model two-players snakes and ladders. The interpretation of x_0 is that the first player wins, and x_1, that the second player wins.

Definition: a hitting time T_x is the first time (possibly infinity) that a state x is reached, i.e. $T_x = \inf\{n : X_n = x\}$.

Idea: solve a bigger problem! What are the hitting probability *for all starting points*, i.e. computing $h(x) = \mathbb{P}_x(T_{x_1} < T_{x_2})$. This allows us to build a recurrence between these problems.

Some easily obtained constraints on h:

1. $h(x_1) = 1$ and $h(x_2) = 0$.
2. $0 \leq h \leq 1$.
3. Exercise (using an argument similar to the one used for the Chapman-Kolmogorov equation):

$$h(x) = \sum_{y \in \mathcal{X}} p(x \to y) h(y).$$

Main result: these constraints can be used to find the numerical value of $h(x)$. More precisely, $h(x)$ is the minimum function satisfying the above three conditions.

9.5 Asymptotic behavior: overview

Assumptions: In the following, we will:

- always assume homogeneity,
- start by assuming finite $\mathcal{X}$, then relax later on.

We will cover two main results, used in different contexts:

1. The law of large number for Markov chains. Informally, that sums converge to a constant:

$$\frac{1}{N} \sum_{n=1}^{N} f(X_n) \xrightarrow{a.s.} c,$$

   where this constant is obtained by an expectation, $c = \mathbb{E}[f(X_\infty)]$, and $\pi(x) = \mathbb{P}(X_\infty = x)$ is called the stationary distribution. This is used for example in the context of Markov chain Monte Carlo (MCMC), covered next week. This only requires a simple condition called irreducibility in our setup.

2. Convergence of marginals:

$$\mathbb{P}(X_n = x) \to \mathbb{P}(X_\infty = x),$$

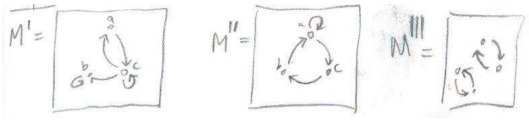
   used for example to determine how many times you need to shuffle a deck of cards. It uses a second condition called aperiodicity in addition to the irreducibility condition mentioned above.

9.6 Law of large number for Markov chains

Definition: a directed path $x \rightsquigarrow y$ in the state diagram of a Markov chain is a list of connected edges $x = x_1 \to x_2 \to \cdots \to x_n = y$ where $(x_i \to x_{i+1}) \in E$.

Definition: a Markov chain is irreducible if there is a directed path between each ordered pair of states.

Example: which of those Markov chains are irreducible, if any?



Solution: the middle one only, M''.

Theorem: if X_n is:

1. Markov,
2. homogeneous,
3. finite,
4. irreducible,

then

$$\frac{1}{N}\sum_{n=1}^{N} f(X_n) \xrightarrow{a.s.} \mathbb{E}[f(X_\infty)],$$

where

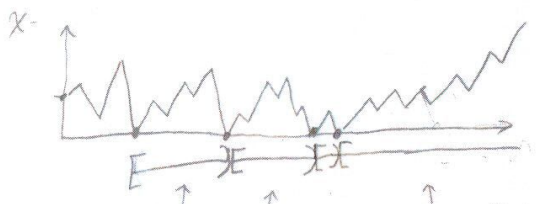
$$\pi(x) = \mathbb{P}(X_\infty = x) = \frac{1}{\mathbb{E}_x T_x}.$$

This is true for any initial distribution μ.

Examples:

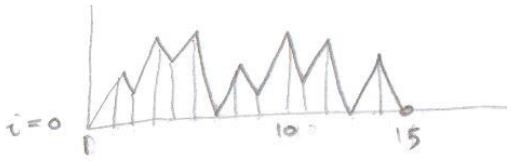
1. Consider the famous board game *Monopoly*. What is the fraction of the rounds where there is a player at one of the squares, say *Park Place*? In this example, $f(x) = \mathbf{1}[x = \text{Park Place}]$.
2. Justification that Markov chain Monte Carlo algorithms can provide arbitrarily good approximations if the user is patient enough, without having to resort to thinning/restarts/burn-in!

Proof idea: i-block, which is a subset of the Markov chain trajectory from one visit to a fixed arbitrary state i to the next visit to i. We will related the average length of these blocks to the inverse of the expected time spent at i. Now these blocks are iid, which will allow us to use the LLN.



Lemma/exercise: $E[T_i] < \infty$ since the chain is finite and irreducible (recall that T_x is a hitting time, defined in the previous section).

Main steps of the proof of LLN:

1. It is enough to show the theorem is true for the test function
$$f(x) = \mathbf{1}[x = i]$$
for some fixed reference state i.
2. Define $N_n = \sum_{j=1}^{n} \mathbf{1}[X_j = i]$, the number of visits to i in the first n steps. We will show that $\frac{1}{n} N_n \to 1/\mathbb{E}_x[T_x]$, almost surely.
3. Define $R(k) = \inf\{n : N_n = k\}$, the time of the k-th return at i.
4. Note: is we let $|B_j|$ denote the length of the j-th i-block, $R(k) = |B_1| + |B_2| + \cdots + |B_{k-1}|$.
5. Hence, by the LLN, $R(k)/(k-1) \to \mathbb{E}_x[T_x]$ almost surely.
6. Note: $R(N_n) \le n \le R(N_n + 1)$. For example, in


we have
$$\underbrace{R(N_n)}_{R(2)=7} \le \underbrace{n}_{10} \le \underbrace{R(N_n + 1)}_{R(3)=13}$$
7. Dividing everything in the above inequalities by N_n, we get $\frac{1}{n} N_n \to 1/\mathbb{E}_x[T_x]$, almost surely.

CLT exercise: a similar argument yield a CLT! Under the same conditions as the previous theorem (note we are still in finite state space, so existence of all moments is guaranteed): there exists a constant c such that
$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} Z,$$
where $Z \sim N(0, c)$.

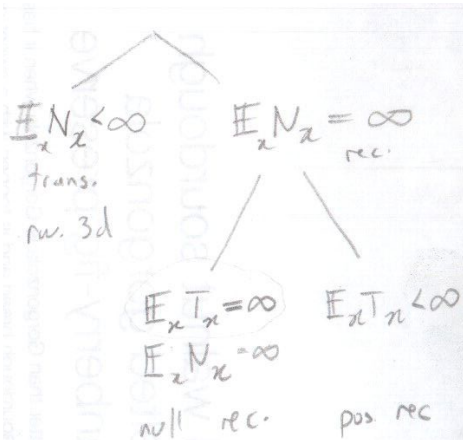
9.7 Extension to countably infinite spaces

The main difficulty is that the fact that $\mathcal{F}$ is countable and X_n is irreducible does not imply that $\mathbb{E}_x T_x < \infty$.

Counter-example: from earlier in the course, a drunk bird might not return home. Recall, using BC 1, if we let $A_n = (X_n = (0,0,0))$, then $\mathbb{P}(A_n \text{ i.o.}) = 0$, implying that $\mathbb{P}_x(T_x = \infty) > 0$, so that $\mathbb{E}_x T_x = \infty$.

Solution: add an assumption. We say that a state is positive recurrent if $\mathbb{E}_x T_x < \infty$. With this additional assumption, the LLN holds in countably infinite spaces.

Note on terminology: why "positive" recurrent? To differentiate from a weaker condition, just called "recurrence", defined by $\mathbb{E}_x N_x = \infty$. Now this is subdivided into two sub-cases, positive recurrent, defined above, and null recurrent, where $\mathbb{E}_x T_x = \infty$ but $\mathbb{E}_x N_x = \infty$. Finally, the opposite of recurrent is transient.



Exercise: find an example which is null recurrent. Hint:

9.8 Convergence of the marginals and coupling

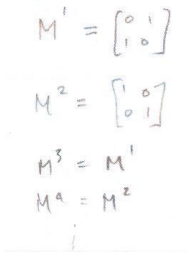
Now, we would like to investigate the convergence of the marginals, which, using our linear algebra, boils down to investigating $\lim_{n\to\infty} M^n$ (note that we get back to the finite case for now).

Problem: just the conditions we used for the LLN are not enough! Counter example:

Note: we have a LLN,

$$\frac{1}{N}\sum_{n=1}^{N}\mathbf{1}[X_n = \text{a}] \xrightarrow{a.s.} \frac{1}{2},$$

but the marginals alternate between two matrices:

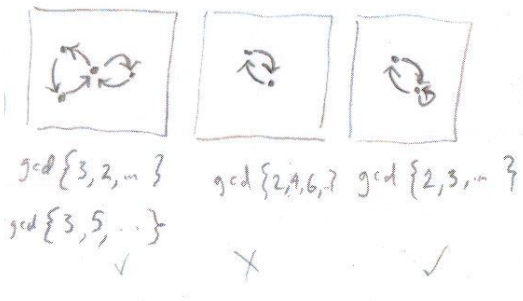


Solution: add an assumption, aperiodicity. To define it, we will need a few definitions.

Definition: the period of $x \in \mathcal{X}$ is defined as $d_x = \gcd\{n : (M^n)_{x,x} > 0\}$.

Definition: a state is aperiodic if $d_x = 1$. A chain is aperiodic if all states are aperiodic.

Examples:



Note that for the first example, we write the gcd relative to two different start states but get the same results. This will always hold in general.

Theorem: if X_n is:

1. Markov,

2. homogeneous,
3. finite,
4. irreducible,
5. aperiodic,

$$\mathbb{P}(X_n = x) \to \mathbb{P}(X_\infty = x),$$

for any initial distribution μ.

Other notation for the result: the above means that $\lim_{n\to\infty} M^n$ exists and is composed of identical rows. Let us denote this limit by L. Let us denote its identical rows by π.

Note: this means that $L = LM$, i.e. $\pi = \pi M$, or

$$\pi(y) = \sum_{x\in\mathcal{X}} \pi(x)p(x \to y).$$

This is called the stationary equation or global balance equation.

Application: debugging of MCMC algorithms (more on this later).

Note: this provide another connection with linear algebra, namely that the stationary equation is an eigenvector of M^T.

Proof idea for the convergence of the marginals: coupling.

1. We build two chains X_n and Y_n. Marginally, both have transition probabilities $p(x \to y)$.
2. $X_0 \sim \pi$, while $Y_0 \sim \mu$.
3. X_n and Y_n make their transitions independently until they meet for the first time, at which point they stay together forever.
4. Note: $X_0 \sim \pi \Rightarrow X_n \sim \pi$ for all n.
5. Let T denote the time the two chains meet: $T = \inf\{n : X_n = Y_n\}$.
6. It is therefore enough to show that the two chain meet "quickly," formally that $\mathbb{P}(T > n) \to 0$. This is where aperiodicity is used! What happens if not irreducible? Consider counter example at the beginning of the section to understand the importance of irreducibility.

7. To conclude the argument, note first that

$$\begin{aligned}
\mathbb{P}(X_n = y, T \le n) &= \sum_{m=1}^{n} \sum_x \mathbb{P}(T = m, X_m = x, X_n = y) \\
&= \sum_{m=1}^{n} \sum_x \mathbb{P}(T = m, X_m = x)\mathbb{P}(X_n = y | T = m, X_m = x) \text{ (chain rule)} \\
&= \sum_{m=1}^{n} \sum_x \mathbb{P}(T = m, Y_m = x)\mathbb{P}(X_n = y | T = m, X_m = x) \text{ (by def. of } T) \\
&= \sum_{m=1}^{n} \sum_x \mathbb{P}(T = m, Y_m = x)\mathbb{P}(Y_n = y | T = m, Y_m = x) \text{ (by homogeneity)} \\
&= \mathbb{P}(Y_n = y, T \le n)
\end{aligned}$$

which implies

$$\begin{aligned}
|\mathbb{P}(X_n = y) - \mathbb{P}(Y_n = y)| &= |(\mathbb{P}(X_n = y, T \le n) + \mathbb{P}(X_n = y, T > n)) \\
&\quad - (\mathbb{P}(Y_n = y, T \le n) + \mathbb{P}(Y_n = y, T > n))| \\
&= |\mathbb{P}(X_n = y, T > n) - \mathbb{P}(Y_n = y, T > n)| \\
&\le \mathbb{P}(X_n = y, T > n) + \mathbb{P}(Y_n = y, T > n),
\end{aligned}$$

hence, summing over y,

$$\begin{aligned}
\sum_y |\mathbb{P}(X_n = y) - \mathbb{P}(Y_n = y)| &\le \sum_y (\mathbb{P}(X_n = y, T > n) + \mathbb{P}(Y_n = y, T > n)) \\
&= 2\mathbb{P}(T > n).
\end{aligned}$$

Definition: the quantity $\frac{1}{2}\sum_y |p(y) - q(y)|$ obtained by rearranging the last inequality in the proof is called the *total variation distance* between distributions p and q.

10 Application: MCMC

10.1 Motivation

Setup: let y denote an observation and x denote an unknown quantity (parameter and/or future or interpolated observation, discrete latent variables, etc).

Two approaches:

1. In practice, at the core of maximum likelihood approaches, a key operation consists in maximizing a likelihood, $x^* = \text{argmax } p_{Y|X}(y|x)$, often via some optimation tools.

2. In practice, at the core of most practical Bayesian approaches, a key operation consists in sampling from a posterior distribution $x^{(i)} \sim p_X(x)p_{Y|X}(y|x)$, often via an approximate sampling method such as MCMC (Markov chain Monte Carlo) or SMC (sequential Monte Carlo).

Cases where the second method is advantageous:

1. Obtaining uncertainty estimates over combinatorial structures. In this case, typical method to get confidence intervals around maximum likelihood estimate either fail (e.g. those based on CLT), or are very inefficient (e.g. the bootstrap).
2. Cases where the maximum of a density is not a good summary. This can arise in situations from partial unidentifiability and stochastic processes for example.

10.2 How to use posterior samples

Let us say we are given $x^{(1)}, x^{(2)}, \cdots \sim p_X(x)p_{Y|X}(y|x)$. How should these be used? There is often something more optimal than say taking the sample with highest posterior (something beginners often resort to).

Bayesian recipe: the Bayesian framework specifies a 3-steps recipe to approach this question.

1. Specify a loss function L over the possible output (decisions/things you are trying to predict). Example: rand loss over clusterings.
2. Compute the posterior distribution. In practice, this is done using an approximate method such via samples $x^{(1)}, x^{(2)}, \ldots$ coming from MCMC.
3. Minimize the posterior expected loss:

$$\text{argmin}_x \mathbb{E}[L(x, X)|Y] \approx \text{argmin} \sum_{i=1}^{N} L(x, x^{(i)}).$$

   Note that we are optimizing, but not a density as in maximum likelihood, rather we are optimizing an integrated loss function. Example/exercise: write the objective function in the case of a rand loss on clusterings.

Pros:

- Statistical efficiency (admissibility, asymptotic efficiency, etc).
- Can be automated via probabilistic programming.
- Combinatorial latent variables supported.
- Correct behavior under partial unidentifiability.

Con: the main con is computational. Many sampling problems have been shown to belong to a provably computationally difficult class of problems called "#P hard problems" (a trickier version of NP hard problems).

10.3 Examples of MCMC algorithms on Ising models

Motivation: computer vision, spatial statistics (see lecture slides for some visualizations).

Basic model: consider a 3x3 grid with each node representing a binary variable. The state space is $\mathcal{X} = +1, -1^{3\cdot 3}$. Notation: if $x \in \mathcal{X}$, we write $x_{i,j}$ for the value of the variable at node at row i and column j in the grid. We write $(i,j) \sim (i',j')$ if two nodes (i,j) and (i',j') are immediate neighbors, i.e. if $|i-i'| + |j-j'| = 1$. Define the following distribution:

$$\pi(x) = \mathbb{P}(X = x) = \frac{1}{Z} \exp\left(\sum_{(i,j)\sim(i',j')} x_{i,j} x_{i',j'} \right). \tag{12}$$

Note: Z can quickly become very hard to compute as the grid gets larger. We have:

$$Z = \sum_{x \in \mathcal{X}} \exp\left(\sum_{(i,j)\sim(i',j')} x_{i,j} x_{i',j'} \right),$$

and just for a 100x100 grid, this means we would have to sum over 2^10000 vectors!

Example of query: what is $\mathbb{P}(X_{1,1} = +1) = \mathbb{E}[\mathbf{1}[X_{1,1} = +1]]$. Here the "test function" is $g(x) = \mathbf{1}[x_{1,1} = +1]$.

Idea:

1. Build/simulate a Markov chain $X^{(1)}, X^{(2)}, \ldots$ where $X^{(t)}$ is a vector $X^{(t)} = (X_1^{(t)}, \ldots, X_9^{(t)})$ taking values in $\mathcal{X}$, and such that the stationary distribution is equal to Equation (12).
2. Use the Law of large numbers for Markov chain!

Challenge: how to create a Markov chain with a prescribed stationary distribution? We will cover two methods (in the analysis, we will reveal that the first is actually a special case of the second):

1. Gibbs sampling.
2. Metropolis-Hastings (MH) algorithms.

10.4 Gibbs sampling

Gibbs algorithm:

1. Initialize the 3x3 grid $x^{(0)}$ arbitrarily.
2. Loop $i = 1, 2, 3, \ldots, N$ (until enough samples are produced):

(a) $x^{(i)} \leftarrow$ copy of $x^{(i-1)}$

(b) Sample one of the 9 variable indices uniformly, $(i^*, j^*) \sim \text{Uni}((1,1),(1,2),\ldots,(3,3))$

(c) Sample a new value $x'_{i^*,j^*} \in \{-1,+1\}$ for $X^{(i)}_{i^*,j^*}$ by sampling from:

$$\mathbb{P}(X_{i^*,j^*} = x'_{i^*,j^*} | X_{i,j} = x^{(i)}_{i,j} \text{ for all } (i,j) \neq (i^*,j^*)). \qquad (13)$$

3. Estimate the expectation(s) of interest using:

$$\frac{1}{N}\sum_{i=1}^{N} g(x^{(i)}).$$

Note: by Bayes rule, Equation (13) is proportional to $\pi(x')\mathbf{1}[x' \in N(x^{(i-1)})]$, where $N(x) = N_{i^*,j^*}(x)$ denotes the set of configurations x' in $\mathcal{X}$ that can be reached from x by changing only variable i^*, j^*. Formally: $N(x) = \{y \in \mathcal{X} : x_{i,j} = y_{i,j} \text{ for all } (i,j) \neq (i^*,j^*)\}$.

Exercise: compute Equation (13) for the Ising example.

Analysis: we will now analyze the behavior of this algorithm as a Markov chain with transitions denoted by p.

- State space: $\mathcal{X}$. Too large to build the transition matrix $M_{x,y} = p(x \to y)$ explicitly! But note that we do not have to if we just want to simulate. Key: each row is sparse. Why?
- Simplification: to start with assume we are always picking a fixed node, say $(2,2)$, in step 2b of the Gibbs algorithm, instead of picking it from a uniform distribution. We will relax this simplifying assumption soon.
- Under this simplification, the form of $p(y \to x)$ is just:

$$p(y \to x) = \frac{\pi(x)\mathbf{1}[x \in N(y)]}{\sum_{x' \in N(y)} \pi(x')}.$$

- Goal: to show that $p(x \to y)$ satisfies the stationary equation, i.e. that if p is as the previous bullet point and π as specified by our target distribution, Equation (12), then we have $\pi(x) = \sum_{y \in \mathcal{X}} \pi(y) p(y \to x)$.

Now we have:

$$\begin{aligned}
\sum_{y \in \mathcal{X}} \pi(y) p(y \to x) &= \sum_{y \in \mathcal{X}} \pi(y) \frac{\pi(x)\mathbf{1}[x \in N(y)]}{\sum_{x' \in N(y)} \pi(x')} \\
&= \pi(x) \sum_{y \in \mathcal{X}} \pi(y) \frac{\mathbf{1}[x \in N(y)]}{\sum_{x' \in N(y)} \pi(x')} \\
&= \pi(x) \frac{\sum_{y \in \mathcal{X}} \pi(y)\mathbf{1}[y \in N(x)]}{\sum_{x' \in N(y)} \pi(x')} \quad (\text{Note that } y \in N(x) \Leftrightarrow x \in N(y)) \\
&= \pi(x).
\end{aligned}$$

10.5 Metropolis-Hastings (MH) algorithms

Limitations of Gibbs: it may be difficult to sample from Equation (13) in certain problems.

MH inputs:

1. A target distribution π that we can evaluate pointwise.
2. A *proposal distribution/transition* $q(x \to y)$ from which we can simulate $q(x \to \cdot)$ and evaluate pointwise.
3. A test function g.

MH algorithm:

1. Initialize:
   (a) $x^{(0)}$ arbitrarily,
   (b) $F \leftarrow 0$
2. Loop $i = 1, 2, 3, \ldots, N$ (until enough samples are produced):
   (a) Propose a new state, $x' \sim q(x^{(i-1)} \to \cdot)$
   (b) Compute:
   $$A(x^{(i-1)} \to x') = \min\left\{1, \frac{\pi(x')q(x' \to x^{(i-1)})}{\pi(x^{(i-1)})q(x^{(i-1)} \to x')}\right\}.$$
   (c) Let $A^{(i)} \sim \text{Bern}(A(x^{(i-1)} \to x'))$
      i. If $A^{(i)} = 1$, then $x^{(i)} \leftarrow x'$
      ii. If $A^{(i)} = 0$, then $x^{(i)} \leftarrow x^{(i-1)}$
   (d) $F \leftarrow F + f(x^{(i)})$
3. Return F/N

Note: the density π always appears in a ratio in the MH algorithm, therefore we do not need to know the normalization constant Z:

$$\frac{\pi(x')}{\pi(x)} = \frac{\gamma(x')/Z}{\gamma(x)/Z} = \frac{\gamma(x')}{\gamma(x)}.$$

Practical note: it is often preferable to compute the numerator and denominator in log-scale and exponentiate only after taking the ratio.

Special cases:

- When q is symmetric (for example, an isotropic normal), the q's cancel out in the ratio.

- When $q(x \to x')$ is independent of x, the algorithm is called an independence chain. Note however that the behavior of the algorithm is still dependent on the previous state because of the accept reject step.

Analysis: we will now analyze the behavior of this algorithm as a Markov chain with transitions denoted by p.

Assume: first that $x \neq y$. What is p? To move from x to y, the chain needs to propose y, and then accept it:

$$p(x \to y) = q(x \to y)A(x \to y).$$

Note: make sure you understand the difference between the proposal q and the Markov chain p used to analyze the algorithm.

Lemma: detailed balance, $\pi(x)p(x \to y) = \pi(y)p(y \to x)$ for all $x, y \in \mathcal{X}$ implies global balance, $\pi(x) = \sum_{y \in \mathcal{X}} p(y \to x)\pi(y)$.

Proof of lemma: sum over y on both sides of the detailed balanced equation.

Proof of MH π-invariance: if $x \neq x'$, we have

$$\begin{aligned}
\pi(x)p(x \to x') &= \pi(x)q(x \to x')A(x \to x') \\
&= \min\{\pi(x)q(x \to x'), \pi(x')q(x' \to x)\} \\
&= \min\{\pi(x')q(x' \to x), \pi(x)q(x \to x')\} \\
&= \pi(x')q(x' \to x)A(x' \to x) \\
&= \pi(x')p(x' \to x).
\end{aligned}$$

Finally, if $x = x'$, the result holds by inspection.

Exercise: show that if q is a conditional distribution of the target distribution, the acceptance ratio is one. Conclude that the Gibbs sampler is a special case of the MH algorithm.

10.6 Irreducibility of MCMC algorithms

Several of the samplers we have defined so far (in particular, the Gibbs sampler) satisfy the global balance equation (i.e. are π-invariant), but they do not have a LLN. Why? Because they are not irreducible. Fortunately, it is easy to restore irreducibility. This is done via combinations of MCMC kernels.

Combination of MCMC kernels. Let us denote a collection of π-invariant kernels by $p_1, p_2, \ldots, p_L$. For example in the Gibbs sampler over a M-by-M grid, we would have one of these kernels for each of the $L = M^2$ nodes. We can combine them using the following methods:

Mixture: where at each step we first pick one of the L kernels and do one MCMC iteration with it. Formally, this create a MCMC kernel given by

$$p_{\text{mix}}(x \to x') = \sum_l \frac{1}{L} p_l(x \to x').$$

Non-uniform distributions over the L kernels could also be used.

Alternation: apply the first kernel, then the second one, then the third one, .., the L-th one, and loop back to the first one. Formally, this create a MCMC kernel given by

$$p_{\text{alt}}(x \to x') = \sum_{x_1 \in \mathcal{X}} \sum_{x_2 \in \mathcal{X}} \cdots \sum_{x_{l-1} \in \mathcal{X}} p_1(x \to x_1) p_2(x_1 \to x_2) \cdots p_L(x_{l-1} \to x').$$

Randomized alternation: first, sample a permutation of $\{1, 2, \ldots, L\}$, second, do one round over all kernels in the order specified by the first step.

Proposition: if each kernel p_l is π-invariant, then the three combinations described above are also π-invariant.

Proof for the mixture:

$$\begin{aligned}
\sum_y \pi(y) p_{\text{mix}}(y \to x) &= \sum_y \pi(y) \sum_l \frac{1}{L} p_l(y \to x) && (14)\\
&= \sum_l \frac{1}{L} \sum_y \pi(y) p_l(y \to x) && (15)\\
&= \sum_l \frac{1}{L} \pi(x) && (16)\\
&= \pi(x). && (17)
\end{aligned}$$

Exercise: prove that the other combination schemes are also π-invariant.

Exercise: conclude that the Ising Gibbs sampler is irreducible and hence that the LLN holds.

References

[1] Imre Bfirfiny and Zoltfin Fiiredi. Computing the volume is difficult. page 8.

[2] Martin Dyer, Alan Frieze, and Ravi Kannan. A Random Polynomial-time Algorithm for Approximating the Volume of Convex Bodies. *J. ACM*, 38(1):1–17, January 1991.

[3] Makoto Matsumoto and Takuji Nishimura. Mersenne Twister: A 623-dimensionally Equidistributed Uniform Pseudo-random Number Generator. *ACM Trans. Model. Comput. Simul.*, 8(1):3–30, January 1998.

[4] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, Cambridge, UK ; New York, 3 edition edition, September 2007.

[5] Rong Zhu and Harry Joe. Negative binomial time series models based on expectation thinning operators. *Journal of Statistical Planning and Inference*, 140(7):1874–1888, July 2010.