PROBABILITY THEORY - PART 1
MEASURE THEORETICAL FRAMEWORK

MANJUNATH KRISHNAPUR

CONTENTS

1. Discrete probability space

"Random experiment" is a non-mathematical term used to describe physical situations with unpredictable outcomes, for instance, "toss a fair coin and observe which side comes up". Can we give a precise mathematical meaning to such a statement? Consider an example.

Example 1

"Draw a random integer from 1 to 100. What is the chance that it is a prime number?"
Mathematically, we just mean the following. Let $\Omega = \{1, 2 \ldots, 100\}$, and for each $\omega \in \Omega$, we set $p_\omega = \frac{1}{100}$. Subsets $A \subseteq \Omega$ are called "events" and for each subset we define $\mathbf{P}(A) = \sum_{\omega \in A} p_\omega$. In particular, if $A = \{2, 3, \ldots, 97\}$ is the set of all prime numbers in Ω, then we get $\mathbf{P}(A) = \frac{1}{4}$.

Whenever there is a random experiment with finitely many or countably many possible outcomes, we can do the same. More precisely, we write Ω for the set of all possible outcomes, and assign the elementary probability p_ω for each $\omega \in \Omega$ (in mathematics, we just assume that these numbers are somehow given. In real life, they will be given by experiments, symmetry considerations etc.). For example,

- Toss a fair coin n times. Here $\Omega = \{0, 1\}^n$ (with the identification that 1 is head and 0 is tail) and $p_\omega = 2^{-n}$ for all $\omega \in \Omega$. If $A = \{\omega \in \Omega : \omega_1 + \ldots + \omega_n = k\}$, then $\mathbf{P}(A) = 2^{-n}\binom{n}{k}$.
- Place r balls in n bins at random. Here Ω is the set of r-tuples with entries from $[n] := \{1, 2, \ldots, n\}$ and $p_\omega = \frac{1}{n^r}$ for each $\omega \in \Omega$.
- Shuffle a deck of n cards. Here Ω is the set of permutations of $[n]$ and $p_\omega = \frac{1}{n!}$ for each $\omega \in \Omega$.
- Throw a biased die n times. Here $\Omega = \{\omega = (i_1, i_2, \ldots, i_n) : 1 \leq i_k \leq 6 \text{ for } k \leq n\}$ is the set of n-tuples with entries from $1, 2, \ldots, 6$. A reasonable assignment of elementary probabilities is $p_\omega = \alpha_{i_1}\alpha_{i_2}\ldots\alpha_{i_n}$ if $\omega = (i_1, \ldots, i_n)$. Here $\alpha_1, \ldots, \alpha_6$ are positive numbers that add up to 1 (and capture the bias in the die).

To conclude, let us make (or recall) a definition.

Definition 1: Discrete probability space

A *discrete probability space* is a pair (Ω, p) where Ω is a finite or countable set and $p : \Omega \to \mathbb{R}_+$ is a function such that $\sum_{\omega \in \Omega} p_\omega = 1$. For a subset $A \subseteq \Omega$, define $\mathbf{P}(A) = \sum_{\omega \in A} p_\omega$.

The only mathematical sophistication needed to understand this definition is the notion of countable sums (convergence, divergence, absolute convergence etc). This we have learned in real analysis class.

Finally, our discussion above may be summarized by saying that the framework of discrete probability spaces captures mathematically the notion of a random experiment with finitely many or countably many possible outcomes. All that is left for a probabilist to do is to take an "interesting" probability space (Ω, p), an "interesting" subset $A \subseteq \Omega$, and actually calculate (or approximately calculate) $\mathbf{P}(A)$. This does not mean it is easy, as the following examples illustrate.

Example 2: Self-avoiding walk

Fix $n \geq 1$ and let Ω be the set of all self-avoiding paths on length n in $\mathbb{Z}^2$ starting from $(0,0)$. That is,

$$\Omega = \{(x_0, \ldots, x_n) : x_0 = (0,0),\ x_i - x_{i-1} \in \{(\pm 1, 0), (0, \pm 1)\} \text{ for } i \leq n \text{ and } x_i \neq x_j \text{ for } i \neq j\}.$$

Let $p_\omega = \frac{1}{\#\Omega}$. One interesting event is $A = \{(x_0, \ldots, x_n) : \|\omega_n\| < n^{0.6}\}$. Far from finding $\mathbf{P}(A)$, it has not been proved whether for large n, the value $\mathbf{P}(A)$ is close to zero or one! If you solve this, click here.

Example 3: Random matrix

Let $\Omega = \{A_{n\times n} : A = (a_{i,j})_{1\leq i,j\leq n},\ a_{i,j} = 0 \text{ or } 1\}$ with $p_\omega = 2^{-n^2}$ for all $\omega \in \Omega$. Let S be the subset of all singular matrices with zero-one entries. What is $\mathbf{P}(S)$? This is a very difficult problem in the field of *random matrix theory*. Partial solutions were achieved by many leading mathematicians before it was solved in 2018 (not an exact solution, but it was shown that asymptotically $\mathbf{P}(A) \approx 2^{-n}$ in an appropriate sense).

Example 4: Percolation

Take the same probability space as in the previous example. Define a path to mean a sequence of indices $(i_1, j_1), \ldots, (i_m, j_m)$ (for some m) such that $i_1 = j_1 = 1$, $i_m = n$ and $(i_{k+1}, j_{k+1}) - (i_k, j_k) \in \{(1,0), (-1,0), (0,1), (0,-1)\}$ for all $1 \leq k \leq m-1$. Let S be the subset of $A_{n\times n}$ for which there is some path for which $a_{i_k, j_k} = 1$ for all k. Finding the probability of S as $n \to \infty$ is an important open problem in a sub-field of probability called *percolation theory* (to be precise, what the answer ought to be is known, proving it is the difficult thing).

Section summary: Random experiments with finite or countably many possible outcomes are adequately modeled mathematically by the notion of a discrete probability space (Ω, p). While calculating probabilities of events may lead to enormous difficulties, the set up itself is mathematically very simple. In short, we know what we are talking about. In the next section we see the difficulties of dealing with uncountable probability spaces.

A guide for the reader familiar with measure theory: If your measure theory knowledge is good, many sections that follow may be skipped or skimmed. Read the axioms of probability in Section 3 and then jump to Section 28 on infinite product spaces, which is where probability ceases to be a branch of general measure theory and becomes a richer subject. There is also material in earlier sections that are usually not covered in measure theory courses. In particular, weak convergence of probability measures in sections 13 and 14 (which will be needed soon), conditional probability in section 24 (which will be needed much later). In addition, to think like a probabilist, one must learn the language of random variables (section 8) and be familiar with the commonly occurring probability distributions (section 12) and have facility with manipulating random variables and their distributions (section 21). Much of the rest is covered in measure theory courses, but it is worth noting certain special features of finite probability spaces (e.g., $L^2 \subseteq L^1$).

2. Uncountable probability spaces?

We want to see how to model random experiments with uncountably many possible outcomes. Start with an example.

Example 5: Break a stick at random

If we idealize the stick to a straight line segment, perhaps a way to make mathematical sense of where it breaks is to pick a point at random from the unit interval. Although it does not sound all that different from picking a number at random from $\{1, 2, \ldots, 100\}$, making sense of this experiment will lead us into very deep waters!

What is so difficult about this? Let us try to imitate what we did before and set $\Omega = [0, 1]$, the set of all possible outcomes. What about probabilities? For example, if $A = [0.1, 0.3]$, then it is clear that we want to say that the probability $\mathbf{P}(A) = 0.2$. Similarly, if $A = \{0.3\}$ or any other singleton, we must assign $\mathbf{P}(A) = 0$.

But then, what is the basis for saying $\mathbf{P}\{[0.1, 0.3]\} = 0.2$? Surely, "$\mathbf{P}\{[0.1, 0.3]\} = \sum_{\omega \in [0.1, 0.3]} p_\omega$" makes no sense?! Since singletons have zero probability, how do we add uncountably many zeros and get a positive number?! Further, what about weird sets, like the set of rational points, the Cantor set, etc? What are their probabilities? You might say

that $\mathbf{P}(A)$ is the length of A for any subset A, but that is not an answer since you have merely replaced the word "probability" by another word "length" (that is, you still have no answer to the question of what is the length of the Cantor set or other weird sets).

Let us mention one other experiment that requires uncountable probability spaces.

Example 6: Toss a fair coin infinitely many times

Here the set of possible outcomes is the uncountable set $\{0,1\}^{\mathbb{N}} = \{0,1\}\times\{0,1\}\times\{0,1\}\times\ldots$. Just as in the case of stick-breaking, there are certain events for which we have no doubt what the probability ought to be. For example, if A is the event that "the first three tosses are heads and the next two are tails", then we have no doubt that the probability must be 2^{-5}.

But again, is this an assumption or a deducible fact? The problem is that any singleton in Ω must have zero probability and summing uncountably many zeros to get 2^{-5} sounds suspicious. Further, there are more complicated events for which it is not clear how to find the probability. For example, events such as "there are infinitely many heads in the sequence" or "after any number of tosses, the number of heads is more than the number of tails" or "for any n, there are at least n heads in the first n^2 tosses", etc.

One can give any number of other examples, for example, "*throw a dart at a dart-board*". But it is enough to keep in mind either the stick breaking example or the coin tossing example. These will turn out to be equivalent. In fact, we shall see later that once we understand one of these examples, we will have understood all uncountable probability spaces! This is true in a precise mathematical sense.

To give a foretaste of how the issues raised in the above examples will be resolved: We shall give up the idea that every subset of the sample space can be assigned probability! Secondly, probabilities of certain (simple) events will be assumed and probabilities of more complicated events will be computed using them. Before coming to this, let us see why such a drastic change of our notions is necessary.

An attempt to fix the issue: Let us stick to the example of drawing a number at random from the interval $[0,1]$ and explain, in a more mathematical manner, the difficulties we run into. We outline a possible approach and see where it runs into difficulties.

Let us *define* the probability of any set $A \subseteq [0,1]$ to be the length of that set. We understand the length of an interval, but what is the length of the set of rational numbers? irrational numbers?

Cantor set? A seemingly reasonable idea is to *define*

$$\mathbf{P}_*(A) = \inf\left\{\sum_{k=1}^{\infty} |I_k| : \text{ each } I_k \text{ is an interval and } \{I_k\} \text{ a countable cover for } A\right\}.$$

and call it the length of A. Then of course, we shall also say that $\mathbf{P}_*(A)$ is the probability of A (in the language of the random experiment, the probability that the chosen random number falls in A). Then perhaps, $\mathbf{P}_*(A)$ should be the probability of A for every subset $A \subseteq [0,1]$. This is at least reasonable in that $\mathbf{P}_*([a,b]) = b - a$ for any $[a,b] \subseteq [0,1]$ (Exercise! This needs proof!). One example of how to compute $\mathbf{P}_*(A)$.

Example 7

Let $A = \mathbf{Q} \cap [0,1]$. Then, we can enumerate A as $\{r_1, r_2, \ldots\}$. Fix $\epsilon > 0$ and let $I_k = [r_k - \epsilon 2^{-k}, r_k + \epsilon 2^{-k}]$ so that $A \subseteq \cup_k I_k$. Further, $\sum_k |I_k| = 2\epsilon$. Since ϵ is arbitrary, this shows that $\mathbf{P}_*(A) = 0$. This is a reasonable answer we might have expected. In fact, for any countable set $A \subseteq [0,1]$, the same argument shows that $\mathbf{P}_*(A) = 0$.

However, we face an unexpected problem. The following fact is not obvious and we do not give a proof now.

Fact 1: Outer measure is not finitely additive

There exists a subset $A \subseteq [0,1]$ such that $\mathbf{P}_*(A) = 1$ and $\mathbf{P}_*(A^c) = 1$.

This fact implies that $\mathbf{P}_*$ cannot be accepted as a reasonable definition of probability, since it violates one of the basic requirements of probability (or of length), that $\mathbf{P}_*(A \cup A^c)$ be equal to $\mathbf{P}_*(A) + \mathbf{P}_*(A^c)$! This approach appears to be doomed to failure.

You may object that our definition of $\mathbf{P}_*$ was arbitrary, and that perhaps a different definition will not run into such absurdities? Before tackling that question, let us be clear about what all properties we want probabilities to satisfy.

We shall certainly want $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$ if A and B are pairwise disjoint subsets of $[0,1]$ (this is called *finite additivity*). But in fact, we shall demand more, that $\mathbf{P}(\cup_{n=1}^{\infty} A_n) = \sum_n \mathbf{P}(A_n)$ if A_n are pairwise disjoint subsets of $[0,1]$. This last requirement is called *countable additivity* and it is not clear why we should ask for it. Honestly, I have no justification to give at this point, except that the accumulated wisdom of mathematicians for about a hundred years has accepted it.

Given these requirements, we run into a serious roadblock.

Result 2

There does not exist[a] any function $f : 2^{[0,1]} \to [0,1]$ such that f is countably additive and $f([a,b]) = b - a$ for all $[a,b] \subseteq [0,1]$.

[a]This result is also not easy to prove. Take it for a fact. For those who are extra curious, here is a bizarre fact: It is possible to find $f : 2^{[0,1]} \to [0,1]$ such that $f(I) = |I|$ for any interval I and such that f is *finitely additive*. However, there does not exist such a finitely additive $f : 2^{[0,1]^3} \to \mathbb{R}$ satisfying $f(I_1 \times I_2 \times I_3) = |I_1| \cdot |I_2| \cdot |I_3|$. In other words, if you want to be a finitely additive probabilist, you may drop countable additivity and happily talk about picking a number at random from an interval, or throw a dart at a board, but not pick a point at random from a cube in three dimensions! Altogether, countable additivity restricts, but leads to a far richer theory within those restrictions.

This means that not only $\mathbf{P}_*$, but any other way you try to define probabilities of subsets of $[0,1]$ (in such a way that $f(I) = |I|$ for intervals), is bound to violate countable additivity and hence, not acceptable to us. This ends our discussion of why we don't know what we are talking about when we said "*draw a number at random from* $[0,1]$". From the next section, we see how this problem can be overcome if we give up our desire to assign probabilities to all subsets.

Section summary: We outlined the various difficulties encountered in giving a mathematical framework for uncountable probability spaces, in particular for the random experiment of breaking a stick at random.

3. Sigma algebras and the axioms of probability

Now we define the setting of probability in abstract and then return to the earlier examples and show how the new framework takes care of the difficulties we discussed.

Definition 2: Probability space

A probability space is a triple $(\Omega, \mathcal{F}, \mathbf{P})$ where

(1) The *sample space* Ω is an arbitrary non-empty set.

(2) The *σ-field* or *σ-algebra* $\mathcal{F}$ is a set of subsets of Ω such that (i) $\emptyset, \Omega \in \mathcal{F}$, (ii) if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$, (iii) if $A_n \in \mathcal{F}$ for $n = 1, 2 \ldots$, then $\cup A_n \in \mathcal{F}$. In words, $\mathcal{F}$ is closed under complementation and under countable unions, and contains the empty set. Elements of $\mathcal{F}$ are called *measurable sets* or *events*.

(3) A *probability measure* is any function $\mathbf{P} : \mathcal{F} \to [0,1]$ is such that if $A_n \in \mathcal{F}$ and are pairwise disjoint, then $\mathbf{P}(\cup A_n) = \sum \mathbf{P}(A_n)$ (countable additivity) and such that $\mathbf{P}(\Omega) = 1$. $\mathbf{P}(A)$ is called *the probability of* A.

By definition, we talk of probabilities only of measurable sets. It is meaningless to ask for the probability of a subset of Ω that is not measurable. Typically, the sigma-algebra will be smaller than the power set of Ω, but large enough to include all sets of interest to us. Restricting the class of sets for which we assign probability is the key idea that will resolve the difficulties we were having with the examples of stick-breaking or infinitely many coin tosses.

The σ-field is closed under many set operations and the usual rules of probability also hold. If one allows $\mathbf{P}$ to take values in $[0, \infty]$ and drops the condition $\mathbf{P}(\Omega) = 1$, then it is just called a *measure*. Measures have the same basic properties as probability measures, but probabilistically crucial concepts of *independence* and *conditional probabilities* (to come later) don't carry over to general measures. Those two concepts are mainly what make probability theory much richer than general measure theory.

Example 8

Let Ω be any non-empty set. Then $\mathcal{F} = 2^{\Omega}$ (collection of all subsets of Ω) is a σ-algebra. The smallest σ-algebra of subsets of Ω is $\mathcal{G} = \{\emptyset, \Omega\}$.

To give an example of a σ-algebra between the two, let $\Omega = \mathbb{R}$ (ar any uncountable set) and define $\mathcal{F}' = \{A \subseteq \Omega : A \text{ or } A^c \text{ is countable}\}$. Check that $\mathcal{F}$ is a σ-algebra. If we define $\mathbf{P}(A) = 0$ if A is countable and $\mathbf{P}(A) = 1$ if A^c is countable, then $\mathbf{P}$ defines a probability measure on $(\Omega, \mathcal{F}')$ (check!).

Some examples of probability spaces. Our new framework better include the old one of discrete probability spaces. Indeed it does, and in that special case, we may also take the sigma-algebra of all subsets of the sample space. This is explained in the following example.

Example 9

Let Ω be a finite or countable set. Let $\mathcal{F}$ be the collection of all subsets of Ω. Then $\mathcal{F}$ is a σ-field. Given a function $p : \Omega \to [0, 1]$ such that $\sum_{\omega \in \Omega} p_\omega = 1$, define $\mathbf{P} : \mathcal{F} \to [0, 1]$ by $\mathbf{P}(A) = \sum_{\omega \in A} p_\omega$. Then, we claim that $\mathbf{P}$ is a probability measure.

To show this we need to show countable additivity. Let $A_1, A_2, \ldots$ be pairwise disjoint subsets of Ω. Countable additivity is the statement that

$$\sum_k \sum_{\omega \in A_k} p_\omega = \sum_{\omega \in \cup_k A_k} p_\omega.$$

If you remember the definition of countable sums, this is an easy exercise (remember that each A_k is countable, possibly finite)[a].

[a] Innumerable times, we shall use without mention the following very important fact: If $a_{i,j} \geq 0$ for $i \geq 1$ and $j \geq 1$, then $\sum_i \sum_j a_{i,j} = \sum_j \sum_i a_{i,j}$ which we simply denote $\sum_{i,j} a_{i,j}$. Further, for any bijection $\sigma : \mathbb{N} \mapsto \mathbb{N} \times \mathbb{N}$,

we have $\sum_{i,j} a_{i,j} = \sum_k a_{\sigma(k)}$. It is highly recommended to brush up basic facts about absolute convergence of series.

More generally, we can have a discrete probability measure inside a 'continuous space'. Such measures also can be defined on the sigma-algebra of all subsets.

Example 10

Let Ω be any set and let $R \subseteq \Omega$ be a countable set. Let $\mathcal{F}$ be the powerset of Ω. Fix non-negative numbers p_x, $x \in R$ that add to 1. Then define $\mathbf{P}(A) = \sum_{x \in R \cap A} p_x$. Then, $\mathbf{P}$ is a probability measure on $\mathcal{F}$ (exercise!).

This means that a discrete measure, say Binomial distribution with parameters n and p, may be considered as a probability measure on $\{0, 1, 2, \ldots, n\}$ or as a probability measure on $\mathbb{R}$ with the power set sigma-algebra. The problem of not being able to define probability for all subsets does not arise in such cases.

A simple exercise about σ-algebras and probability measures.

Exercise 1

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space.

(1) $\mathcal{F}$ is closed under finite and countable unions, intersections, differences, symmetric differences. Also $\Omega \in \mathcal{F}$.

(2) If $A_n \in \mathcal{F}$, then

$$\limsup A_n := \{\omega : \omega \text{ belongs to infinitely many } A_n\},$$

$$\liminf A_n := \{\omega : \omega \text{ belongs to all but finitely many } A_n\}$$

are also in $\mathcal{F}$. In particular, if A_n increases or decreases to A, then $A \in \mathcal{F}$.

(3) $\mathbf{P}(\emptyset) = 0$, $\mathbf{P}(\Omega) = 1$. For any $A, B \in \mathcal{F}$ we have $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$. If $A_n \in \mathcal{F}$, then $\mathbf{P}(\cup A_n) \leq \sum \mathbf{P}(A_n)$.

(4) If $A_n \in \mathcal{F}$ and A_n increases (decreases) to A, the $\mathbf{P}(A_n)$ increases (decreases) to $\mathbf{P}(A)$.

Generated σ-algebras: In the most interesting cases, one cannot explicitly say what the elements of $\mathcal{F}$ are, but only require that it is rich enough that it contains sets of interest to us. We make a simple observation.

Exercise 2

Let $\mathcal{F}_\alpha$, $\alpha \in I$ be a colllection of σ-algebras of subsets of Ω (here I is an arbitrary index set). Then let $\mathcal{F} = \bigcap_{\alpha \in I} \mathcal{F}_\alpha$. Show that $\mathcal{F}$ is a σ-algebra.

In particular, if S is a collection of subsets of Ω, then show that there is a smallest σ-algebra $\mathcal{F}$ containing S (this means that $\mathcal{F}$ is a σ-algebra and any σ-algebra containing S also contains $\mathcal{F}$). We say that $\mathcal{F}$ is generated by S and write $\mathcal{F} = \sigma(S)$.

Caution: Note that the only definition of $\sigma(S)$ is that it is the smallest sigma algebra containing S. It is *not* true that it is the collection of all countable unions (or countable unions of countable intersections or countable unions of countable intersections of countable unions ...) elements of S. As an analogy, consider a vector space V and a subset of vectors S. The subspace generated by S has two equivalent definitions: (1) It is the smallest subspace of V that contains S and (2) it is the set of all finite linear combinations of elements of S. The second definition may be called internal, while the first is external. For generated sigma algebras, we have no internal definition, only an external one.

Stick-breaking example: In the new language that we have introduced, let us revisit the question of making mathematical sense of stick-breaking. Let $\Omega = [0,1]$ and let S be the collection of all intervals. To be precise let us take all right-closed, left-open intervals $(a,b]$, with $0 \le a < b \le 1$ as well as intervals $[0,b]$, $b \le 1$ (alternate description: take all intervals of the form $(u,v] \cap [0,1]$ where $u < v$ are real numbers). If we are trying to make precise the notion of '*drawing a number at random from* $[0,1]$', then we would want $\mathbf{P}(a,b] = b - a$ and $\mathbf{P}[0,b] = b$. The precise mathematical questions can now be formulated as follows.

Question 1

(1) Let $\mathcal{G}$ be the σ-algebra of all subsets of $[0,1]$. Is there a probability measure $\mathbf{P}$ on $\mathcal{G}$ such that $\mathbf{P}(a,b] = b - a$ and $\mathbf{P}[0,b] = b$ for all $0 \le a < b \le 1$?

(2) Let $\mathcal{F} = \sigma(S)$ be the Borel σ-algebra of $[0,1]$. Is a probability measure $\mathbf{P}$ on $\mathcal{F}$ satisfying $\mathbf{P}(a,b] = b - a$ and $\mathbf{P}[0,b] = b$ for all $0 \le a < b \le 1$?

The answer to the first question is 'No' (this was stated as Result **??**), which is why we need the notion of σ-fields, and the answer to the second question is 'Yes', which is why probabilists still have their jobs. Neither answer is obvious, but we shall answer them in coming lectures.

Coin-tossing example: Let $\Omega = \{0,1\}^{\mathbb{N}} = \{\omega = (\omega_1, \omega_2, \ldots) : \omega_i \in \{0,1\}\}$. Let S be the collection of all subsets of Ω that depend on only finitely many co-ordinates (such sets are called cylinders). More precisely, a cylinder set is of the form $A = \{\omega : \omega_{k_1} = \epsilon_1, \ldots \omega_{k_n} = \epsilon_n\}$ for some given $n \geq 1$, $k_1 < k_2 < \ldots < k_n$ and $\epsilon_i \in \{0,1\}$ for $i \leq n$.

What are we talking about? If we want to make precise the notion of '*toss a coin infinitely many times*', then clearly Ω is the sample space to look at. It is also desirable that elements of S be in the σ-field as we should be able to ask questions such as 'what is the chance that the fifth, seventh and thirtieth tosses are head, tail and head respectively' which is precisely asking for the probability of a cylinder set.

If we are '*tossing a coin with probability p of turning up Head*', then for a cylinder set $A = \{\omega : \omega_{k_1} = \epsilon_1, \ldots \omega_{k_n} = \epsilon_n\}$, it is clear that we would like to assign $\mathbf{P}(A) = \prod_{i=1}^n p^{\epsilon_i} q^{1-\epsilon_i}$ where $q = 1 - p$. So the mathematical questions are: (i) If we take $\mathcal{F}$ to be the σ-field of all subsets of Ω, does there exist a probability measure $\mathbf{P}$ on $\mathcal{F}$ such that for cylinder sets $\mathbf{P}(A)$ is as previously specified. (ii) If the answer to (i) is 'No', is there a probability measure $\mathbf{P}$ on $\mathcal{F}$ such that for a cylinder set A as above, $\mathbf{P}(A) = \prod_{i=1}^n p^{\epsilon_i} q^{1-\epsilon_i}$?

Again, the answers are 'No' and 'Yes', respectively.

The σ-fields in these two examples can be captured under a common definition.

Definition 3: Borel sigma algebra

Let (X, d) be a metric space. The σ-field $\mathcal{B}$ generated by all open balls in X is called the Borel sigma-algebra of X.

First consider $[0,1]$ or $\mathbb{R}$. Let $S = \{(a,b]\} \cup \{[0,b]\}$ and let $T = \{(a,b)\} \cup \{[0,b)\} \cup \{(a,1]\}$. We could also simply write $S = \{(a,b] \cap [0,1] : a < b \in \mathbb{R}\}$ and $T = \{(a,b) \cap [0,1] : a < b \in \mathbb{R}\}$. Let the sigma-fields generated by S and T be denoted $\mathcal{F}$ (see example above) and $\mathcal{B}$ (Borel σ-field), respectively. Since $(a,b] = \cap_n (a, b + \frac{1}{n})$ and $[0,b] = \cap_n [0, b + \frac{1}{n})$, it follows that $S \subseteq \mathcal{B}$ and hence $\mathcal{F} \subseteq \mathcal{B}$. Similarly, since $(a,b) = \cup_n (a, b - \frac{1}{n}]$ and $[0,b) = \cup_n [0, b - \frac{1}{n}]$, it follows that $T \subseteq \mathcal{F}$ and hence $\mathcal{B} \subseteq \mathcal{F}$. In conclusion, $\mathcal{F} = \mathcal{B}$.

In the countable product space $\Omega = \{0,1\}^{\mathbb{N}}$ or more generally $\Omega = X^{\mathbb{N}}$, the topology is the one generated by all sets of the form $U_1 \times \ldots \times U_n \times X \times X \times \ldots$ where U_i are open sets in X. Clearly each of these sets is a cylinder set. Conversely, each cylinder set is an open set. Hence $\mathcal{G} = \mathcal{B}$. More generally, if $\Omega = X^{\mathbb{N}}$, then cylinders are sets of the form $A = \{\omega \in \Omega : \omega_{k_i} \in B_i, i \leq n\}$ for some $n \geq 1$ and $k_i \in \mathbb{N}$ and some Borel subsets B_i of X. It is easy to see that the σ-field generated by cylinder sets is exactly the Borel σ-field.

We shall usually work with Borel σ-algebras of various metric spaces, as this σ-algebra is rich enough to contain almost all sets we might be interested in. If you are not convinced, try finding

a subset of $[0,1]$ that is not a Borel set (it is quite a non-trivial exercise!). Here are some easy exercises.

Exercise 3

On $\mathbb{R}^d$, show that each of the following classes of sets generates the Borel σ-algebra of $\mathbb{R}^d$ (particularly think about the case $n = 1$).

(1) The collection of all open balls.

(2) The collection of all closed balls.

(3) The collection of all closed rectangles of the form $[a_1, b_1] \times \ldots \times [a_n, b_n]$ for $a_i < b_i$.

(4) Same as before, but let the rectangles be left-open and right-closed, i.e, sets of the form $(a_1, b_1] \times \ldots \times (a_n, b_n]$ for $a_i < b_i$.

4. The 'standard trick' of measure theory!

While we care about sigma fields only, there are smaller sub-classes that are useful in elucidating the proofs. Here we define some of these.

Definition 4

Let S be a collection of subsets of Ω. We say that S is a

(1) **π-system** if $A, B \in S \implies A \cap B \in S$.

(2) **λ-system** if (a) $\Omega \in S$, (b) $A, B \in S$ and $A \subseteq B \implies B \backslash A \in S$, (c) $A_n \uparrow A$ and $A_n \in S \implies A \in S$.

(3) **Algebra** if (a) $\emptyset, \Omega \in S$, (b) $A \in S \implies A^c \in S$, (c) $A, B \in S \implies A \cup B \in S$.

(4) **Monotone class** if (a) $A_n \in S$ and $A_n \uparrow A \implies A \in S$ and (b) $A_n \in S$ and $A_n \downarrow A \implies A \in S$. Recall that $A_n \uparrow A$ means that $A_1 \subseteq A_2 \subseteq \ldots$ and $\cup_n A_n = A$ and $A_n \downarrow A$ means that $A_1 \supseteq A_2 \supseteq \ldots$ and $\cap_n A_n = A$.

(5) **σ-algebra** if (a) $\emptyset, \Omega \in S$, (b) $A \in S \implies A^c \in S$, (c) $A_n \in S \implies \cup A_n \in S$.

We have included the last one again for comparision. Clearly a sigma algebra is a π-system, a λ-system, a monotone class and an algebra. The difference between algebras and σ-algebras is just that the latter is closed under countable unions while the former is closed only under finite unions. As with σ-algebras, arbitrary intersections of algebras/λ-systems/π-systems are again algebras/λ-systems/π-systems and hence one can talk of the algebra generated by a collection of subsets or a λ-system generated by a collection of subsets etc.

Example 11

The table below exhibits some examples.

Ω	S ($\pi-$system)	$\mathcal{A}(S)$ (algebra generated by S)	$\sigma(S)$
$(0,1]$	$\{(a,b] : 0 < a \leq b \leq 1\}$	$\{\cup_{k=1}^{N} I_k : I_k \in S$ are pairwise disjoint$\}$	$\mathcal{B}(0,1]$
$[0,1]$	$\{(a,b] \cap [0,1] : a \leq b\}$	$\{\cup_{k=1}^{N} R_k : R_k \in S$ are pairwise disjoint$\}$	$\mathcal{B}[0,1]$
$\mathbb{R}^d$	$\{\prod_{i=1}^{d}(a_i,b_i] : a_i \leq b_i\}$	$\{\cup_{k=1}^{N} R_k : R_k \in S$ are pairwise disjoint$\}$	$\mathcal{B}_{\mathbb{R}^d}$
$\{0,1\}^{\mathbb{N}}$	collection of all cylinder sets	finite disjoint unions of cylinders	$\mathcal{B}(\{0,1\}^{\mathbb{N}})$

Often, as in these examples, sets in a π-system and in the algebra generated by the π-system can be described explicitly, but not so the sets in the generated σ-algebra. This point, that a Borel set is not easily expressed by a countable number of operations on intervals, is at the heart of the non-triviality of the subject. Now we present two useful lemmas that allow us to say things about a sigma algebra even when its elements are "out of touch". The spirit of both lemmas is the same, and in many occasions they may be used interchangeably.

Lemma 3: Sierpinski-Dynkin π-λ theorem

Let Ω be a set and let $\mathcal{F}$ be a set of subsets of Ω.

(1) $\mathcal{F}$ is a σ-algebra if and only if it is a π-system as well as a λ-system.

(2) If S is a π-system, then $\lambda(S) = \sigma(S)$.

Lemma 4: Monotone class theorem

Let Ω be a set and let S be a collection of subsets of Ω. If S is an algebra, then the monotone class generated by S is a sigma-algeba. That is, $\mathcal{M}(S) = \sigma(S)$.

Proof of the π-λ theorem. (1) One way is clear. For the other way, suppose $\mathcal{F}$ is a π-system as well as a λ-system. Then, $\Omega \in \mathcal{F}$ and if $A \in \mathcal{F}$, then $A^c = \Omega \backslash A \in \mathcal{F}$. If $A_n \in \mathcal{F}$, then the finite unions $B_n := \cup_{k=1}^{n} A_k = (\cap_{k=1}^{n} A_k^c)^c$ belong to $\mathcal{F}$ (for intersections use that $\mathcal{F}$ is a π-system). The countable union $\cup A_n$ is the increasing limit of B_n and hence belongs to $\mathcal{F}$ by the λ-property.

(2) By the first part, it suffices to show that $\mathcal{F} := \lambda(S)$ is a π-system, that is, we only need show that if $A, B \in \mathcal{F}$, then $A \cap B \in \mathcal{F}$. This is the tricky part of the proof!

Fix $A \in S$ and let $\mathcal{F}_A := \{B \in \mathcal{F} : B \cap A \in \mathcal{F}\}$. S is a π-system, hence $\mathcal{F}_A \supset S$. We claim that $\mathcal{F}_A$ is a λ-system. Clearly, $\Omega \in \mathcal{F}_A$. If $B, C \in \mathcal{F}_A$ and $B \subseteq C$, then $(C \backslash B) \cap A = (C \cap A) \backslash (B \cap A) \in \mathcal{F}$ because $\mathcal{F}$ is a λ-system containing $C \cap A$ and $B \cap A$. Thus $(C \backslash B) \in \mathcal{F}_A$. Lastly, if $B_n \in \mathcal{F}_A$ and $B_n \uparrow B$, then $B_n \cap A \in \mathcal{F}_A$ and $B_n \cap A \uparrow B \cap A$. Thus $B \in \mathcal{F}_A$. This

means that $\mathcal{F}_A$ is a λ-system containing S and hence $\mathcal{F}_A \supset \mathcal{F}$. In other words, $A \cap B \in \mathcal{F}$ for all $A \in S$ and all $B \in \mathcal{F}$.

Now fix any $A \in \mathcal{F}$. And again define $\mathcal{F}_A := \{B \in \mathcal{F} : B \cap A \in \mathcal{F}\}$. Because of what we have already shown, $\mathcal{F}_A \supset S$. Show by the same arguments that $\mathcal{F}_A$ is a λ-system and conclude that $\mathcal{F}_A = \mathcal{F}$ for all $A \in \mathcal{F}$. This is another way of saying that $\mathcal{F}$ is a π-system. ■

Exercise 4: Monotone class theorem

Follow similar steps and prove the monotone class theorem. Note that you only need to show that $\mathcal{M}(S)$ is a sigma algebra.

As an application, we prove a certain uniqueness of extension of measures. The question is this: if two probability measures on $(\mathbb{R}, \mathcal{B})$ agree on all intervals, then are the same? It is tempting to say yes, since intervals generate the Borel sigma-algebra. But this reasoning is false as the following example shows.

Example 12

Let $\Omega = \{1, 2, 3, 4\}$ and let $S = \{\{1, 2\}, \{2, 3\}, \{3, 4\}\}$. Then it is easy to see that $\sigma(S) = 2^\Omega$ (the power set). Now define two probability measures μ, ν on Ω by setting $\mu_i = \frac{1}{4}$ for all i while $\nu_1 = \nu_3 = \frac{1}{2}$ while $\nu_2 = \nu_4 = 0$. Then $\mu(A) = \frac{1}{2} = \nu(A)$ for all $A \in S$, although $\mu \neq \nu$ on $\sigma(S)$.

It may be worth recalling here our earlier analogy with vector spaces and generated subspaces. If two linear transformations agree on a collection of vectors, then they agree on the subspace generated by those vectors. This is trivial since every vector in the generated subspace is a linear combination of vectors in the given collection. The example above shows that the lack of an "internal definition" for the generated sigma algebra is not only an inconvenience, but the analogous statement is even false!

Here is a positive result in this direction.

Lemma 5

Let S be a π-system of subsets of Ω and let $\mathcal{F} = \sigma(S)$. If $\mathbf{P}$ and $\mathbf{Q}$ are two probability measures on $\mathcal{F}$ such that $\mathbf{P}(A) = \mathbf{Q}(A)$ for all $A \in S$, then $\mathbf{P}(A) = \mathbf{Q}(A)$ for all $A \in \mathcal{F}$.

Proof. Let $\mathcal{G} = \{A \in \mathcal{F} : \mathbf{P}(A) = \mathbf{Q}(A)\}$. By the hypothesis $\mathcal{G} \supseteq S$. We claim that $\mathcal{G}$ is a λ-system. Clearly, $\Omega \in \mathcal{G}$. If $A, B \in \mathcal{G}$ and $A \supseteq B$, then $\mathbf{P}(A \backslash B) = \mathbf{P}(A) - \mathbf{P}(B) = \mathbf{Q}(A) - \mathbf{Q}(B) = \mathbf{Q}(A \backslash B)$, implying that $A \backslash B \in \mathcal{G}$. Lastly, if $A_n \in \mathcal{G}$ and $A_n \uparrow A$, then $\mathbf{P}(A) = \lim_{n\to\infty} \mathbf{P}(A_n) =$

$\lim_{n\to\infty} \mathbf{Q}(A_n) = \mathbf{Q}(A)$ (this follows from countable additivity of measures). Thus $\mathcal{G} \supseteq \lambda(S)$ which is equal to $\sigma(S)$ by the π-λ theorem. Thus $\mathbf{P} = \mathbf{Q}$ on $\mathcal{F}$. ■

Remark 1

To emphasize the point again, typically, our σ-algebras (eg., the Borel σ-algebra) are defined as being generated by a given collection of sets (eg., left-open right-closed intervals). While the sets in the algebra generated by this collection can often be expressed explicitly in terms of the sets in the collection (eg., finite unions of pairwise disjoint left-open right-closed intervals), the sets in the σ-algebra are more intangible[a] (most emphatically Borel sets are not always countable unions of intervals!). Hence, to show that a property holds for all elements of the σ-algebra, we simply consider the collection of all sets having that property, and show that the collection is a σ-algebra. In doing that, we may find it easier to show that it is a λ-system or that it is a monotone class (containing an appropriate π-system or an algebra).

[a] There is a way to express them, using transfinite induction. But let us ignore that approach and stick to the definition which simply says that it is the smallest σ-algebra containing...

5. Lebesgue measure

Theorem 6: Existence and uniqueness of Lebesgue measure

There exists a unique Borel measure λ on $[0,1]$ such that $\lambda(I) = |I|$ for any interval I.

Note that $S = \{(a,b] \cap [0,1]\}$ is a π-system that generates $\mathcal{B}$. Therefore by Lemma 5, uniqueness follows. Existence is all we need to show.

There are several steps in the proof of existence. We outline the big steps and leave some routine checks to the reader. In this proof, Ω will denote $[0,1]$.

Step 1 - Definition of the outer measure λ_*: Define $\lambda_*(A)$ for any subset by

$$\lambda_*(A) = \inf\left\{\sum |I_k| : \text{ each } I_k \text{ is an open interval and } \{I_k\} \text{ a countable cover for } A\right\}.$$

(In the definition, we could have used closed intervals or left-open right-closed intervals to cover A. It is easy to see that the value of $\lambda_*(A)$ remains unchanged.)

Check that λ_* has the following properties. (1) $0 \le \lambda_*(A) \le 1$ is a well-defined for every subset $A \subseteq \Omega$, (2) $\lambda_*(A \cup B) \le \lambda_*(A) + \lambda_*(B)$ for any $A, B \subseteq \Omega$, (3) $\lambda_*(\Omega) = 1$. Two remarks.

(1) For the last property, try the more general Exercise **??** below.

(2) Clearly, from finite subadditivity, we also get countable subadditivity $\lambda_*(\cup A_n) \leq \sum \lambda_*(A_n)$. The difference from a measure is that equality might not hold, even if there are finitely many sets and they are pairwise disjoint.

(3) The three properties above constitute the definition of what is called an *outer measure*.

Exercise 5

Show that $\lambda_*(a, b] = b - a$ if $0 < a \leq b \leq 1$.

Step-2 - The σ-field on which λ_* will be shown to be a measure: Let λ_* be an outer measure on a set Ω. Caratheodary's brilliant definition is to set

$$\mathcal{F} := \{A \subseteq \Omega : \lambda_*(E) = \lambda_*(A \cap E) + \lambda_*(A^c \cap E) \text{ for any } E\}.$$

Note that subadditivity implies $\lambda_*(E) \leq \lambda_*(A \cap E) + \lambda_*(A^c \cap E)$ for any E for any A. The non-trivial requirement is the inequality in the reverse direction.

Claim 7

$\mathcal{F}$ is a sigma algebra and λ_* restricted to $\mathcal{F}$ is a probability measure.

Proof. It is clear that $\emptyset, \Omega \in \mathcal{F}$ and $A \in \mathcal{F}$ implies $A^c \in \mathcal{F}$. Next, suppose $A, B \in \mathcal{F}$. Then for any E,

$$\begin{aligned}
\lambda_*(E) &= \lambda_*(E \cap A) + \lambda_*(E \cap A^c) \\
&= \lambda_*(E \cap A \cap B) + \lambda_*(E \cap A \cap B^c) + \lambda_*(E \cap A^c) \\
&\geq \lambda_*(E \cap A \cap B) + \lambda_*(E \cap (A \cap B)^c)
\end{aligned}$$

where the last inequality holds by subadditivity of λ_* and $(E \cap A \cap B^c) \cup (E \cap A^c) = E \cap (A \cap B)^c$. Hence $\mathcal{F}$ is a π-system.

As $A \cup B = (A^c \cap B^c)^c$, it also follows that $\mathcal{F}$ is an algebra. For future use, note that $\lambda_*(A \cup B) = \lambda_*(A) + \lambda_*(B)$ if A, B are disjoint sets in $\mathcal{F}$. To see this apply the definition of $A \in \mathcal{F}$ with $E = A \cup B$.

To show that $\mathcal{F}$ is a σ-algebra, by the $\pi - \lambda$ theorem, it suffices to show that $\mathcal{F}$ is a λ-system. Suppose $A, B \in \mathcal{F}$ and $A \supseteq B$. Then

$$\begin{aligned}
\lambda_*(E) &= \lambda_*(E \cap B^c) + \lambda_*(E \cap B) \\
&= \lambda_*(E \cap B^c \cap A) + \lambda_*(E \cap B^c \cap A^c) + \lambda_*(E \cap B) \\
&\geq \lambda_*(E \cap (A \backslash B)) + \lambda_*(E \cap (A \backslash B)^c).
\end{aligned}$$

Thus $A \backslash B \in \mathcal{F}$. It remains to show closure under increasing limits,

Suppose $A_n \in \mathcal{F}$ and $A_n \uparrow A$. Then $\lambda_*(A) \geq \lambda_*(A_n) = \sum_{k=1}^{n} \lambda_*(A_k \backslash A_{k-1})$ by finite additivity of λ_*. Hence $\lambda_*(A) \geq \sum \lambda_*(A_k \backslash A_{k-1})$. The other way inequality follows by subadditivity of λ_* and

we get $\lambda_*(A) = \sum \lambda_*(A_k \backslash A_{k-1})$. Then for any E we get

$$\begin{aligned}\lambda_*(E) &= \lambda_*(E \cap A_n) + \lambda_*(E \cap A_n^c) \\ &\geq \lambda_*(E \cap A_n) + \lambda_*(E \cap A^c) \\ &= \sum_{k=1}^{n} \lambda_*(E \cap (A_k \backslash A_{k-1})) + \lambda_*(E \cap A^c).\end{aligned}$$

The last equality follows by finite additivity of λ_* on $\mathcal{F}$ (which we showed above). Let $n \to \infty$ and use subadditivity to see that

$$\begin{aligned}\lambda_*(E) &\geq \sum_{k=1}^{\infty} \lambda_*(E \cap (A_k \backslash A_{k-1})) + \lambda_*(E \cap A^c) \\ &\geq \lambda_*(E \cap A) + \lambda_*(E \cap A^c).\end{aligned}$$

Thus, $A \in \mathcal{F}$ and it follows that $\mathcal{F}$ is a λ-system too and hence a σ-algebra.

Lastly, if $A_n \in \mathcal{F}$ are pairwise disjoint with union A, then $\lambda_*(A) \geq \lambda_*(\cup_{k=1}^n A_k) = \sum_{k=1}^n \lambda_*(A_k) \to \sum_k \lambda_*(A_k)$ while the other way inequality follows by subadditivity of λ_* and we see that $\lambda_*|_{\mathcal{F}}$ is a measure. ■

Step-3 - $\mathcal{F}$ is large enough: We want to show that $\mathcal{F}$ contains all Borel sets. Since $\mathcal{F}$ is already shown to be a σ-algebra, and the Borel σ-algebra is generated by left-open, right-closed intervals, the following claim is all we need.

Claim 8

Let $A = (a, b] \subseteq [0, 1]$. Then $A \in \mathcal{F}$.

Proof. For any $E \subseteq [0, 1]$, let $\{I_n\}$ be an open cover such that $\lambda_*(E) + \epsilon \geq \sum |I_n|$. Then, note that $\{I_n \cap (a, b)\}$ and $\{I_n \cap [a, b]^c\}$ are open covers for $A \cap E$ and $A^c \cap E$, respectively ($I_n \cap [a, b]^c$ may be a union of two intervals, but that does not change anything essential). It is also clear that $|I_n| = |I_n \cap (a, b)| + |I_n \cap (a, b)^c|$. Hence we get

$$\lambda_*(E) + \epsilon \geq \sum |I_n \cap (a, b)| + \sum |I_n \cap (a, b)^c| \geq \lambda_*(A \cap E) + \lambda_*(A^c \cap E).$$

This holds for any $\epsilon > 0$ and hence $\lambda_*(E) \geq \lambda_*(A \cap E) + \lambda_*(A^c \cap E)$. By subadditivity we always have $\lambda_*(E) \leq \lambda_*(A \cap E) + \lambda_*(A^c \cap E)$. Thus we see that $A \in \mathcal{F}$. ■

Conclusion: We have obtained a σ-algebra $\mathcal{F}$ that is larger than the $\mathcal{B}$ and such that μ_* is a probability measure when restricted to $\mathcal{F}$. Hence μ_* is also a probability measure when restricted to $\mathcal{B}$. The proof of Theorem 6 is complete.

6. Further remarks on the Lebesgue measure, its construction and life in general

6.1. **Borel and Lebesgue σ-algebras.** We have $\mathcal{B} \subseteq \mathcal{F} \subseteq 2^{[0,1]}$ (recall that 2^{Ω} denotes the powerset of Ω). Are these containments strict? How much smaller is $\mathcal{B}$ compared to $\mathcal{F}$?

Elements of $\mathcal{F}$ are called Lebesgue measurable sets. Below we show that there is a subset of $[0,1]$ that is not Lebesgue-measurable. Now let us consider the relationship between $\mathcal{B}$ and $\mathcal{F}$. This is explained more in to homework problems, but we make short remarks.

(1) The cardinality of $\mathcal{B}$ is the same as that of $\mathbb{R}$ while the cardinality of $\mathcal{F}$ is the same as that of $2^{\mathbb{R}}$. Thus, in this sense, $\mathcal{F}$ is much larger than $\mathcal{B}$.

(2) For probability, the difference is less serious. For any set $A \in \mathcal{F}$, there are two sets $B, C \in \mathcal{B}$ such that $B \subseteq A \subseteq C$ and such that $\mu(B) = \mu(C)$. In other words, the only new sets that enter into $\mathcal{F}$ are those that can be sandwiched between Borel sets of equal measure. The weird thing about the Borel σ-algebra is that even if $A_1 \subseteq A_2$, $A_2 \in \mathcal{B}$ and $\mu(A_2) = 0$, it may happen that A_1 is not in $\mathcal{B}$ (and hence we cannot write $\mu(A_1) = 0$). The Lebesgue σ-algebra does not have this issue (it is called the *completion* of the Borel σ-algebra with respect to Lebesgue measure). Henceforth, if needed, we write $\bar{\mathcal{B}}$ for the Lebesgue σ-algebra.

Nevertheless, we shall put all our probability measures on the Borel σ-algebra. The reason is that completion of a σ-algebra (see Homework 1), although harmless, depends on the measure with respect to which we complete. Since we often need to consider many probability measures at the same time, it is more convenient to work with the Borel sigma algebra.

In the next section we show that $\mathcal{F}$ is strictly smaller than the power set, i.e., there exists sets that are not Lebesgue measurable. Thus, both the containments in $\mathcal{B} \subseteq \mathcal{F} \subseteq 2^{[0,1]}$ are strict.

6.2. **Sigma-algebras are necessary.** We have already mentioned that there is no *translation invariant* probability measure on all subsets of $[0,1]$ (non-measurable sets are shown in the next section). Hence, we had to restrict to a smaller σ-algebra ($\mathcal{B}$ or $\bar{\mathcal{B}}$). If we do not require translation invariance for the extended measure, the question becomes more difficult.

Note that there do exist probability measures on the σ-algebra of all subsets of $[0,1]$, so one cannot say that there are no measures on all subsets. For example, define $\mathbf{Q}(A) = 1$ if $0.4 \in A$ and $\mathbf{Q}(A) = 0$ otherwise. Then $\mathbf{Q}$ is a probability measure on the space of all subsets of $[0,1]$. $\mathbf{Q}$ is a discrete probability measure in hiding! If we exclude such measures, then it is true that some subsets have to be omitted to define a probability measure You may find the proof for the following general theorem in Billingsley, p. 46 (uses *axiom of choice* and *continuum hypothesis*).

Fact 9

There is no probability measure on the σ-algebra of all subsets of $[0,1]$ that gives zero probability to singletons.

Say that x is an atom of $\mathbf{P}$ if $\mathbf{P}(\{x\}) > 0$ and that $\mathbf{P}$ is purely atomic if $\sum_{\text{atoms}} \mathbf{P}(\{x\}) = 1$. The above fact says that if $\mathbf{P}$ is defined on the σ-algebra of all subsets of $[0,1]$, then $\mathbf{P}$ must be have atoms. It is not hard to see that in fact $\mathbf{P}$ must be purely atomic. To see this let $\mathbf{Q}(A) = \mathbf{P}(A) - \sum_{x \in A} \mathbf{P}(\{x\})$. Then $\mathbf{Q}$ is a non-negative measure without atoms. If $\mathbf{Q}$ is not identically zero, then with $c = \mathbf{Q}([0,1])^{-1}$, we see that $c\mathbf{Q}$ is a probability measure without atoms, and defined on all subsets of $[0,1]$, contradicting the stated fact. This last manipulation is often useful and shows that we can write any probability measure as a convex combination of a purely atomic probability measure and a completely nonatomic probability measure

Remark 2: Importance of sigma algebras

The discussion so far shows that σ-algebras cannot be avoided. In measure theory, they are pretty much a necessary evil. However, in probability theory, σ-algebras have much greater significance as place holders of information. Even if Lebesgue measure were to exist on all subsets, probabilisits would have had to invent the concept of σ-algebras! These cryptic remarks are not meant to be understood yet, but we shall have occasion to explain it later in the course.

6.3. **Finitely additive measures.** If we relax countable additivity, strange things happen. For example, there does exist a *translation invariant* ($\mu(A + x) = \mu(A)$ for all $A \subseteq [0,1]$, $x \in [0,1]$, in particular, $\mu(I) = |I|$) *finitely additive* ($\mu(A \cup B) = \mu(A) + \mu(B)$ for all A, B disjoint) probability measure defined on all subsets of $[0,1]$! In higher dimensions, even this fails, as shown by the mind-boggling

Theorem 10: Banach-Tarski paradox

he unit ball in $\mathbb{R}^3$ can be divided into finitely many (five, in fact) disjoint pieces and rearranged (only translating and rotating each piece) into a ball of twice the original radius!!

In some sense, this makes finitely additive measure less attractive to us as a framework for probability theory. In the finitely additive framework, we can break a stick at random (and ask for probability that the break-point is any subset of $[0,1]$) but we cannot break three sticks and ask the same question (that the break points belong to an arbitrary subset of $[0,1]^3$)! The objection is perhaps not entirely acceptable to everyone. In any case, it is a good policy in life to accept

countably additive measures as the right framework for probability, but keep in mind that life can change and finitely additive measures may become more important in some future contexts.

6.4. **How general is the construction of Lebesgue measure?** The construction of Lebesgue measure can be made into a general procedure for constructing interesting measures, starting from measures of some rich enough class of sets. The steps are as follows.

(1) Given an algebra $\mathcal{A}$ (in this case finite unions of $(a, b]$), and a *countably additive p.m* $\mathbf{P}$ on $\mathcal{A}$, define an outer measure $\mathbf{P}_*$ on all subsets by taking infimum over countable covers by sets in $\mathcal{A}$.

(2) Then define $\mathcal{F}$ exactly as above, and prove that $\mathcal{F} \supset \mathcal{A}$ is a σ-algebra and $\mathbf{P}_*$ is a probability measure on $\mathcal{A}$.

(3) Show that $\mathbf{P}_* = \mathbf{P}$ on $\mathcal{A}$.

Proofs are quite the same. Except, in $[0, 1]$ we started with λ defined on a π-system S rather than an algebra. But in this case the generated algebra consists precisely of disjoint unions of sets in S, and hence we knew how to define λ on $\mathcal{A}(S)$. When can we start with $\mathbf{P}$ defined on a π-system? The crucial point in $[0, 1]$ was that for any $A \in S$, one can write A^c as a finite union of sets in S. In such cases (which includes examples from the previous lecture) the generated algebra is precisely the set of disjoint finite unions of sets in S. If that is the case, we define $\mathbf{P}$ on $\mathcal{A}(S)$ in the natural manner and then proceed to step one above.

Following the general procedure outlined above, one can construct the following probability measures.

(1) A probability measure on $([0, 1]^d, \mathcal{B})$ such that $\mathbf{P}([a_1, b_1] \times \ldots \times [a_d, b_d]) = \prod_{k=1}^{d} (b_k - a_k)$ for all cubes contained in $[0, 1]^d$. This is the d-dimensional Lebesgue measure.

(2) A probability measure on $\{0, 1\}^{\mathbb{N}}$ such that for any cylinder set $A = \{\omega : \omega_{k_j} = \epsilon_j,\ j = 1, \ldots, n\}$ (any $n \geq 1$ and $k_j \in \mathbb{N}$ and $\epsilon_j \in \{0, 1\}$) we have (for a fixed $p \in [0, 1]$ and $q = 1 - p$)

$$\mathbf{P}(A) = \prod_{j=1}^{n} p^{\epsilon_j} q^{1-\epsilon_j}.$$

(3) Let $F : \mathbb{R} \to [0, 1]$ be a non-decreasing, right-continuous function such that $\lim_{x\to\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$ (such a function is called a *cumulative distribution function* or CDF in short). Then, there exists a unique probability measure μ on $(R, \mathcal{B}_{\mathbb{R}})$ such that $\mu(a, b] = F(b) - F(a)$ for all $a < b$.

BUT we want to de-emphasize this approach. Instead, we want to emphasize that Lebesgue measure is the only measure that needs to be constructed. We can take the existence of Lebesgue

measure as a black-box, and use it to construct all other probability measures that we need. This includes the above three classes of examples and every probability measure of interest to probabilists. That is subject of the next few sections.

7. Non-measurable sets

Sigma-algebras would not be necessary in measure theory if all subsets of $[0,1]$ were Lebesgue measurable. In this section, we show that non-measurable sets do exist. Let $\bar{\mathcal{B}}$ denote the Lebesgue σ-algebra.

We change the setting a little bit. Let us consider the sample space $[0,1)$ which is a group under addition modulo 1. By $\mathcal{B}$ and $\bar{\mathcal{B}}$ we mean the Borel and Lebesgue σ-agebras of $[0,1)$ and let λ be the Lebesgue measure on $[0,1)$. You may either think of repeating the whole procedure of construction with $[0,1)$ in place of $[0,1]$ or more simply, note that $\mathcal{B}_{[0,1)} = \{A \cap [0,1) : A \in \mathcal{B}_{[0,1]}\}$ and similarly for $\bar{\mathcal{B}}_{[0,1)}$. Further, λ is the restriction to $[0,1)$. We shall need the following 'translation invariance property' of λ on $\bar{\mathcal{B}}$.

Exercise 6

Show that for any $A \in \bar{\mathcal{B}}$ and $x \in [0,1]$ that $A + x \in \bar{\mathcal{B}}$ and that $\lambda(A+x) = \lambda(A)$.

To clarify the notation, for any $A \subseteq [0,1]$ and any $x \in [0,1]$, $A + x := \{y + x \pmod 1 : y \in A\}$. For example, $[0.4, 0.9] + 0.2 = [0, 0.1] \cup [0.6, 1)$.

First construction of a non-measurable set: Now we construct a subset $A \subseteq [0,1]$ and countably (infinitely) many $x_k \in [0,1]$ such that the sets $A + x_k$ are pairwise disjoint and $\cup_k (A + x_k)$ is the whole of $[0,1]$. Then, if A were in $\bar{\mathcal{B}}$, by the exercise $A + x_k$ would have the same probability as A. But $\sum \lambda(A + x_k)$ must be equal to $\lambda([0,1]) = 1$, which is impossible! Hence $A \notin \bar{\mathcal{B}}$.

How to construct such a set A and $\{x_k\}$? Define an equivalence relation on $[0,1]$ by $x \sim y$ if $x - y \in \mathbb{Q}$ (check that this is indeed an equivalence relation). Then, $[0,1]$ splits into pairwise disjoint equivalence classes whose union is the whole of $[0,1]$. Invoke *axiom of choice* to get a set A that has exactly one element from each equivalence class. Consider $A + r$, $r \in \mathbb{Q} \cap [0,1)$. If $A + r$ and $A + s$ intersect then we get an $x \in [0,1]$ such that $x = y + r = z + s \pmod 1$ for some $y, z \in A$. This implies that $y - z = r - s \pmod 1$ and hence that $y \sim z$. So we must have $y = z$ (as A has only one element from each equivalence class) and that forces $r = s$ (why?). Thus the sets $A + r$ are pairwise disjoint as r varies over $\mathbb{Q} \cap [0,1)$. Further given $x \in [0,1]$, there is a $y \in A$ belonging to the equivalence class of x. Therefore $x \in A + r$ where $r = x - y$ (if $y \leq x$) or $r = x - y + 1$ (if $x < y$). Thus we have constructed the set A whose countably many translates $A + r$, $r \in \mathbb{Q} \cap [0,1)$ are pairwise disjoint! Thus, A is a subset of $[0,1]$ that is not Lebesgue measurable.

Remark 3

In mathematical jargon, if $G = \mathbb{Q} \cap [0,1)$ is a subgroup of $[0,1)$, and A is a set which contains exactly one representative of each coset of this subgroup. Then, for each $x \in A$ the set $x + G$ is the coset containing x and hence $\bigsqcup_{r \in G} (A + r) = [0,1)$. As G is countable, by the argument outlined above, it follows that A cannot be Lebesgue measurable.

A second construction showing that λ_* is not finitely additive: Now we want to construct $B \subseteq [0,1)$ such that $\lambda_*(B) = 1$ and $\lambda_*(B^c) = 1$. Then of course, B cannot be measurable (why?). But this example is stronger than the previous one as it shows that on the power-set of $[0,1)$, the outer measure fails finite additivity, not just countable additivity.

I would have liked to take $R \subseteq \mathbb{Q} \cap [0,1)$ and set $B = \bigsqcup_{r \in R} (A + r)$ so that $B^c \bigsqcup_{r \in R^c} (A + r)$ with A as in the previous construction. We already know that $\lambda_*(A) > 0$ (any set of outer measure 0 is measurable), so the hope would be that if both R and R^c are infinite (or suitably large), then $\lambda_*(B) = 1$ and $\lambda_*(B^c) = 1$. But I was not able to prove that any subset R works. If you can show that, I would be very interested to know!

One of the difficulties is that ideally one would like to divide $\mathbb{Q} \cap [0,1)$ into two "equal" subsets R and R^c. For example, if we could find R such that R^c is a translate of R (i.e., $R^c = r_0 + R$), then B^c would be a translate of B and hence they would have the same outer measure (that does not complete the proof, but I am trying to motivate what we do next). But we cannot find such as set R because $\mathbb{Q} \cap [0,1)$ does not have subgroups of finite index!

What is the way out? Let consider a different group $G = \{n\alpha : n \in \mathbb{Z}\}$ (here and below, we are working within $[0,1)$, hence $n\alpha$ always means $n\alpha$ (mod 1) etc.), where α is an irrational number in $[0,1)$, eg., $1/\sqrt{2}$.

Exercise 7

Show that (1) $n\alpha \neq m\alpha$ for all $m \neq n$, (2) G is a subgroup of $[0,1)$ that is isomorphic to $\mathbb{Z}$, (3) G is dense in $[0,1)$.

Let $H = \{2n\alpha : n \in \mathbb{Z}\}$. Then H is a subgroup of G and it has only two cosets, H and $H' := H + \alpha$. If you have done the previous exercise, you will easily see that H and H' are both dense in $[0,1)$.

By the axiom of choice, chose a subset $A \subseteq [0,1)$ that has exactly one representative in each coset of G (as a subgroup of $[0,1)$). Define $B = A + H = \{a + h : a \in A, h \in H\}$. Then $B^c = A + H' = A + H + \alpha$.

We claim that $(B-B)\cap H'=\emptyset$. Indeed, any element of $B-B$ is of the form $a+h-a'-h'$ where $a,a'\in A$ and $h,h'\in H$. If $a=a'$, then this element is in H and hence not in H'. If $a\neq a'$, by construction of A we know that $a-a'\notin G$. But $h-h'\in G$ and hence $a+h-a-h'$ is not in G and hence not in H' either. This proves the claim.

Note that $B^c-B^c=B-B$ (an element of B^c-B^c is of the form $(a+h+\alpha)-(a'+h'+\alpha)=(a+h)-(a'+h')$). Therefore, we also have $(B^c-B^c)\cap H'=\emptyset$.

To proceed, we need the following important fact.

Lemma 11: Steinhaus' lemma

et $A\subseteq[0,1)$ be a measurable subset of positive Lebesgue measure. Then $A-A$ contains an interval around 0. More explicitly, there is some $\delta>0$ such that $(1-\delta,1)\cup[0,\delta)\subseteq A-A$.

Now we claim that $\lambda_*(B^c)=1$. If not, suppose $\lambda_*(B^c)<1-\epsilon$. By definition of outer measure, find intervals I_k such that $\cup I_k\supseteq B^c$ and $\sum_k|I_k|<1-\epsilon$. Then consider $C:=\cap I_k^c=(\cup I_k)^c$. Obviously C is a Lebesgue measurable set, $C\subseteq B$, and $\lambda(C)=1-\lambda(\cup I_k)\geq 1-\sum_k\lambda(I_k)>\epsilon$. Thus $C-C$ contains an interval by Steinhaus' lemma. Since $B\supseteq C$, we also see that $B-B$ contains an interval. But this contradicts the fact that H' is dense, since we have shown that $(B-B)\cap H'=\emptyset$. Thus we must have $\lambda_*(B^c)=1$. An identical argument (since B^c-B^c is also disjoint from H') shows that $\lambda_*(B)=1$.

It only remains to prove Steinhaus' lemma.

Proof of Steinhaus' lemma. By definition of outer measure, there is a covering of A by countably many intervals I_k such that $\lambda(A)\geq 0.9\sum_k|I_k|$. But $\lambda(A)\leq\sum_k\lambda(A\cap I_k)$. Hence, there is at least one k for which $\lambda(A\cap I_k)\geq 0.9\lambda(I_k)>0$. For simplicity, write I for this I_k and let $A'=A\cap I$.

Fix $x\in\mathbb{R}$ and note that

$$\begin{aligned}\lambda(A'\cap(A'+x))&=\lambda(A')+\lambda(A'+x)-\lambda(A'\cup(A'+x))\\&\geq 2\lambda(A')-\lambda(I\cup(I+x))\\&\geq 1.8|I|-(|I|+|x|)\end{aligned}$$

which is positive for $|x|<\delta:=0.8|I|$. In particular, for $|x|<\delta$, we have $A'\cup(A'+x)\neq\emptyset$. Rephrasing this, we see that $x\in A'-A'\subseteq A-A$. ■

Both Steinhaus' lemma and the following fact (whose proof was implicit in the above proof) are very useful tools in measure theory.

Fact 12

Let $A\subseteq\mathbb{R}$ be a measurable subset with $\lambda(A)>0$. Then, for any $\epsilon>0$, there is an interval I (depending on ϵ) such that $\lambda(A\cap I)\geq(1-\epsilon)\lambda(I)$.

Remark 4

There is a theorem of Solovay to the effect that the axiom of choice is necessary to show the existence of a non-measurable set (as an aside, we should perhaps not have used the word 'construct' given that we invoke the axiom of choice). We see that it was used in both constructions above. In the problem set, another construction due to Sierpinski is outlined, and that also uses the axiom of choice.

8. Random variables

Definition 5: Random variables

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and let $(X, \mathcal{G})$ be a set with a σ-algebra. A function $T : \Omega \mapsto X$ is called a random variable (or measurable function) if $T^{-1}A \in \mathcal{F}$ for any $A \in \mathcal{G}$. Here $T^{-1}(A) := \{\omega \in \Omega : T(\omega) \in A\}$ for any $A \subseteq X$.
Generally, we take X to be a metric space and $\mathcal{G} = \mathcal{B}_X$, in which case we say that T is an X-valued random variable.

Important cases: When $X = \mathbb{R}$ we just say T is a "random variable" and when $X = \mathbb{R}^d$ we say T is a "random vector". When $X = C[0,1]$ with its Borel sigma algebra (under the sup-norm metric $d(f,g) = \max\{|f(t) - g(t)| : t \in [0,1]\}$), T is called a "stochastic process" or a "random function". When X is itself the space of all locally finite countable subsets of $\mathbb{R}^d$ (with Borel sigma algebra in an appropriate metric which I do not want to mention now) , we call T a "point process". In genetics or population biology one looks at genealogies, and then we have tree-valued random variables, in the study of random networks, we have random variables taking values in the set of all finite graphs etc, etc.

Remark 5

Some remarks.

(1) Let Ω_1, Ω_2 be two non-empty sets and let $T : \Omega_1 \to \Omega_2$ be a function.
   (a) Suppose we fix a σ-algebra $\mathcal{G}$ on Ω_2. Then, the "pull-back" $\{T^{-1}A : A \in \mathcal{G}\}$ is the smallest σ-algebra on Ω_1 w.r.t. which T is measurable (if we fix $\mathcal{G}$ on Ω_2) . We write $\sigma(T)$ for this σ algebra. In older notation, it is $\sigma(\mathcal{S})$ where $\mathcal{S} = \{T^{-1}A : A \in \mathcal{G}\}$.
   (b) Suppose we fix a σ-algebra $\mathcal{F}$ on Ω_1. The "push-forward" $\{A \subseteq \Omega_2 : T^{-1}A \in \mathcal{F}\}$ is the largest σ-algebra on Ω_2 w.r.t. which T is measurable (if we fix $\mathcal{F}$ on Ω_1).

   That they are σ-algebras is a consequence of the fact that $T^{-1}(A)^c = T^{-1}(A^c)$ and $T^{-1}(\cup A_n) = \cup_n T^{-1}(A_n)$ (Caution! It is generally false that $T(A^c) = T(A)^c$).

(2) Let $\mathcal{F}$ and $\mathcal{G}$ be σ-algebras on Ω_1 and Ω_2, respectively. If $\mathcal{S}$ generates $\mathcal{G}$, i.e., $\sigma(\mathcal{S}) = \mathcal{G}$, then to check that T is measurable, it suffices to check that $T^{-1}A \in \mathcal{F}$ for any $A \in \mathcal{S}$. In particular, $T : \Omega \to \mathbb{R}$ is measurable if and only if $T^{-1}(-\infty, x] \in \mathcal{F}$ for any $x \in \mathbb{R}$.

(3) It is convenient to allow random variables to take the values $\pm\infty$. In other words, when we say random variable, we mean $T : \Omega \to \bar{\mathbb{R}}$ where the set of extended real numbers $\mathbb{R} \cup \{+\infty, -\infty\}$ is a metric space with the metric $d(x, y) = |\tan^{-1}(x) - \tan^{-1}(y)|$ with $\tan^{-1} : \bar{\mathbb{R}} \mapsto [-\frac{\pi}{2}, \frac{\pi}{2}]$. The metric is not important (there are many metrics we can choose from), what matters are the open sets. Open sets in $\bar{\mathbb{R}}$ include open subsets of $\mathbb{R}$ as well as sets of the form $(a, +\infty]$ and $[-\infty, a)$. Similarly, random vectors will be allowed to take values in $(\bar{\mathbb{R}})^d$.

(4) If $A \subseteq \Omega$, then the *indicator function* of A. $\mathbf{1}_A : \Omega \to \mathbb{R}$ is defined by $\mathbf{1}_A(\omega) = 1$ if $\omega \in A$ and $\mathbf{1}_A(\omega) = 0$ if $\omega \in A^c$. If $\mathcal{F}$ is a σ-algebra on Ω, observe that $\mathbf{1}_A$ is a random variable if and only if $A \in \mathcal{F}$.

Example 13

Consider $([0, 1], \mathcal{B})$. Any continuous function $T : [0, 1] \to \mathbb{R}$ is a random variable. This is because $T^{-1}(\text{open}) = \text{open}$ and open sets generate $\mathcal{B}(\mathbb{R})$.

Here is an interesting point (a curiosity since we have said that we shall work with $\mathcal{B}$, not $\overline{\mathcal{B}}$).

Exercise 8

If we endow $\mathbb{R}$ with the Lebesgue sigma algebra $\overline{\mathcal{B}}$, show that there are continuous functions from $\mathbb{R}$ to itself that are not measurable!

Random variables are closed under many common operations. As an illustration, suppose $X, Y : \Omega \to \mathbb{R}$ are random variables and let $Z = X + Y$. We want to show that Z is a random variable. Indeed,

$$\begin{aligned} Z^{-1}(-\infty, t) &= \{\omega : Z(\omega) < t\} \\ &= \{\omega : X(\omega) < s \text{ and } Y(\omega) < t - s \text{ for some } s\} \\ &= \bigcup_{s \in \mathbb{Q}} (X^{-1}(-\infty, s)) \cap (Y^{-1}(-\infty, t - s)) \end{aligned}$$

which is in the σ-algebra, being from by countably many intersections and unions of sets in the σ-algebra. A small point to note is that if we work with $Z^{-1}(-\infty, t]$, then the proof will have to

be modified a little (if $t = 0$, $X = -Y = \sqrt{2}$, then we cannot find $s \in \mathbb{Q}$ such that $X \leq s$ and $Y \leq t - s$).

Note the importance of taking $s \in \mathbb{Q}$ to get countable unions. Similarly or more easily, solve the exercises below. Remember to allow $\pm\infty$ as possible values.

Exercise 9

Show that $T : \mathbb{R} \to \mathbb{R}$ is measurable if it is any of the following. (1) lower semicontinuous function, (2) right continuous function, (3) step function, (4) non-decreasing function.

Exercise 10

Let $(\Omega, \mathcal{F})$ be a measurable space.

(1) If T_1, T_2 are random vectors on Ω, and $a, b \in \mathbb{R}$, then $aT_1 + bT_2$ is a random vector.

(2) If $T = (T_1, \ldots, T_d)$ where $T_i : \Omega \to \bar{\mathbb{R}}$, then T is a random vector if and only if each T_i is a random variable.

(3) Supremum (or infimum) of a countable family of random variables is a random variable.

(4) The $\limsup$ (or $\liminf$) of a countable sequence of random variables is a random variable.

Push forward of a measure: If $T : \Omega_1 \to \Omega_2$ is a random variable, and $\mathbf{P}$ is a probability measure on $(\Omega_1, \mathcal{F}_1)$, then defining $\mathbf{Q}(A) = \mathbf{P}(T^{-1}A)$, we get a p.m $\mathbf{Q}$, on $(\Omega_2, \mathcal{F}_2)$. $\mathbf{Q}$, often denoted $\mathbf{P}T^{-1}$ is called the push-forward of $\mathbf{P}$ under T.

The reason why $\mathbf{Q}$ is a measure is that if A_n are pairwise disjoint, then $T^{-1}A_n$ are pairwise disjoint. However, note that if B_n are pairwise disjoint in Ω_1, then $T(B_n)$ are in general not disjoint. This is why there is no "pull-back measure" in general (unless T is one-one, in which case the pull-back is just the push-forward under T^{-1}!)

When $(\Omega_2, \mathcal{F}_2) = (\mathbb{R}, \mathcal{B})$, the push forward (a Borel p.m on $\mathbb{R}$) is called the *distribution* of the r.v. T. If $T = (T_1, \ldots, T_d)$ is a random vector, then the pushforward, a Borel probability measure on $\mathbb{R}^d$ is called the distribution of T or as the *joint distribution* of $T_1, \ldots, T_d$. Note that all probabilistic questions about a random variable can be answered by knowing its distribution. The original sample space is irrelevant. If X and Y are random variables having the same distribution, by definition, $\mathbf{P}\{X \in A\} = \mathbf{P}\{Y \in A\}$ for any A in the range-space.

Remark 6

Random variables in "real situations". Consider a real-life random experiment, for example, a male-female pair have a child. What is the sample space? For simplicity let us think of genetics as determining everything. Then, the male and female have their DNAs which are two strings of four alphabets, i.e., they are of the form $(A, T, T, G, C, C, C, \ldots, G)$ whose lengths are about 10^9. These two strings are given (nothing random about them, let us assume).

The child to be born can (in principle) have any possible DNA where each element of the string comes from the father or the mother. This large collection of strings is the sample space (its cardinality is less than 2^{10^9}, but perhaps 2^{10^8} or so). The actual probability distribution on these strings is very complicated and no one can write it down explicitly, but for simplicity you may think that it is uniform (equal probability for all possible strings).

Even after the child is born, we do not know ω, i.e., we do not observe the DNA of the child. What we observe are various functions of the DNA string, such as "colour of the eye", "weight at birth", etc. These observations/measurements are random variables. We can also plot the height or weight of the offspring from birth to death - that gives us a random function.

Similarly, in any realistic random experiment, the outcome we see is not ω, but values of a few random variables $X(\omega), Y(\omega) \ldots$. Our questions are also about random variables. For example, we may ask, "what is the probability that the weight of the child after one month is less than 3 kg.?". As remarked earlier, all we need is the distribution of the random variable $X :=$ weight of the child after one month.

9. Borel Probability measures on Euclidean spaces

Given a metric space X, let $\mathcal{P}(X)$ denote the space of all Borel probability measures on X. We want to understand $\mathcal{P}(\mathbb{R}^d)$ for $d \geq 1$.

So far, the only probability measure that we know is the Lebesgue measure λ on $[0, 1]$. Can we at least construct a few more examples. Indeed, if $T : [0, 1] \to \mathbb{R}^d$ is any Borel-measurable function, then $\lambda \circ T^{-1}$ gives a Borel probability measure on $\mathbb{R}^d$. This gives us a large collection of examples of probability measures. The surprising result that we shall see is that there are no others!

Theorem 13

Let μ be a Borel probability measure on $\mathbb{R}^d$. Then, there exists a Borel function $T : [0, 1] \to \mathbb{R}^d$ such that $\mu = \lambda \circ T^{-1}$.

One nice thing about this is that we understand functions better than measures, and the above theorem says that every Borel probability measure can be got using a Borel function. However, the map T is not unique. Indeed, consider $T, T' : [0,1] \to \mathbb{R}$ defined by $T(x) = x$ and $T'(x) = 1 - x$. Then the push-forward of λ under both T and T' is λ itself. It would be nicer to associate to each probability measure, a unique function. This is done by the useful idea of a distribution function.

Definition 6

Let μ be a Borel probability measure on $\mathbb{R}^d$. Define its *cumulative distribution function* (abbreviated as CDF, also simply called distribution function) $F_\mu : \mathbb{R}^d \to [0,1]$ by $F_\mu(x) = \mu(R_x)$ where $R_x := (-\infty, x_1] \times \ldots \times (-\infty, x_d]$ for $x = (x_1, \ldots, x_d)$. In $d = 1$ in particular, $F_\mu(t) = \mu(-\infty, t]$.

Distribution functions have three key properties.

(1) F_μ is non-decreasing in each co-ordinate.

(2) F_μ is right continuous in each co-ordinate.

(3) If $\min_i x_i \to -\infty$, then $F_\mu(x) \to 0$. If $\min_i x_i \to +\infty$, then $F_\mu(x) \to 1$.

The first property is obvious because $R_x \subseteq R_y$ if $x_i \leq y_i$ for each $i \leq d$. For the second property, we note that if $x^{(n)} = (x_1^{(n)}, \ldots, x_d^{(n)})$ and $x_i^{(n)} \downarrow y_i$, then the sets $R_{x^{(n)}}$ decrease to R_y. Hence, by the properties of measures, $\mu(R_{x^{(n)}}) \downarrow \mu(R_y)$ which is precisely the right-continuity of F_μ. For the third listed property, we note that if $\min_i x_i \downarrow -\infty$ (respectively, $\min_i x_i \uparrow +\infty$), then $R_{x^{(n)}}$ decreases to the empty set (respectively, increases to $\mathbb{R}^d$). Again, the behaviour of measures under increasing and decreasing limits (which is equivalent to countable additivity) implies the stated properties.

We caution the reader on two common mistakes.

(1) F_μ is not left-continuous in general. Taking $d = 1$ for simplicity of notation, note that if $t_n \uparrow t$, then $(-\infty, t_n]$ increases to $(-\infty, t)$, not to $(-\infty, t]$. Hence, left-continuity may not hold (examples below show it too).

(2) $F_\mu(x) \to 0$ if just one of the x_is goes to $-\infty$ but to have $F_\mu(x) \to 1$, we need (in general) all x_is to go to $+\infty$. In $d = 2$, for example, if $x_1 \uparrow \infty$ and x_2 stays fixed, then $R_x \uparrow \mathbb{R} \times (-\infty, x_2]$ and not to $\mathbb{R}^2$.

As we have only a few examples of probability measures so far, we give two examples.

Example 14

Let μ be the Lebesgue measure on $[0,1]$. Then,

$$F_\mu(t) = \begin{cases} 0 & \text{if } t \leq 0, \\ t & \text{if } 0 \leq t \leq 1, \\ 1 & \text{if } t \geq 1. \end{cases}$$

Example 15

Let $\mu = \delta_0$, which is the probability measure defined by $\delta_0(A) = 1$ if $A \ni 0$ and $\delta_0(A) = 0$ if $A \not\ni 0$. Then, we see that

$$F_{\delta_0}(t) = \begin{cases} 0 & \text{if } t < 0, \\ 1 & \text{if } t \geq 0. \end{cases}$$

This is an example where left-continuity fails at one point. More generally, consider a discrete measure $\mu = \sum_{k=1}^n q_k \delta_{a_k}$ for some real numbers $a_1 < \ldots < a_n$ and for some non-negative numbers q_i such that $q_1 + \ldots + q_n = 1$. Its distribution function is given by

$$F_\mu(t) = \begin{cases} 0 & \text{if } t < a_1, \\ q_1 + \ldots + q_i & \text{if } a_i \leq t < a_{i+1}, \\ 1 & \text{if } t \geq a_n. \end{cases}$$

It fails left-continuity at $a_1, \ldots, a_n$.

Exercise 11

Define the probability measure $\delta_{(0,0)}$ on $\mathbb{R}^2$. Write its distribution function. Do the same for $\frac{1}{4}(\delta_{(0,0)} + \delta_{(0,1)} + \delta_{(1,0)} + \delta_{(1,1)})$.

Now we come to the second theorem which shows that distribution functions are in one-one correspondence with Borel probability measures.

Theorem 14

Suppose $F : \mathbb{R}^d \to [0,1]$ is nondecreasing in each co-ordinate, right-continuous in each co-ordinate, and satisfies $\lim F(x) = 0$ if $\min_i x_i \to -\infty$ and $\lim F(x) = 1$ if $\min_i x_i \to +\infty$. Then, there exists a unique Borel probability measure μ on $\mathbb{R}^d$ such that $F_\mu = F$.

The uniqueness part is easy. If μ and ν are two Borel probability measures on $\mathbb{R}^d$ having the same distribution function, then $\mu(R_x) = \nu(R_x)$ for all $x \in \mathbb{R}^d$. But the collection $\mathcal{S} := \{R_x : x \in \mathbb{R}^d\}$ is a π-system that generates the Borel σ-algebra. Hence $\mu = \nu$.

The difficult part is existence of a mmeasure μ. In the next two sections, we prove Theorem 13 and Theorem 14, first for $d = 1$ and then for general d.

10. The case of one-dimension

For $d = 1$, we prove Theorem 13 and Theorem 14 simultaneously (I am unable to find such a proof for higher dimensions[1]).

Suppose $F : \mathbb{R} \to [0, 1]$ satisfying the assumptions of Theorem 14 is given. Define $T : (0, 1) \to \mathbb{R}$ by

$$T(u) := \inf\{x : F(x) \geq u\}.$$

Since we restrict to $(0, 1)$, it follows that T is well-defined (since $F(x)$ converges to 0 and 1 at $-\infty$ and $+\infty$). Further, T is non-decreasing and left continuous. In particular, it is Borel-measurable. Hence, $\mu := \lambda \circ T^{-1}$ is a well-defined Borel probability measure on $\mathbb{R}$. We claim that μ has distribution function F.

What is T? When F is strictly increasing and continuous, T is just the inverse of F. In general, it is a sort of generalized inverse in the sense that $T(u) \leq x$ if and only if $F(x) \geq u$. Hence,

$$\begin{aligned}\lambda \circ T^{-1}(-\infty, x] &= \lambda\{u \in (0, 1) : T(u) \leq x\} \\ &= \lambda\{u \in (0, 1) : u \leq F(x)\} \\ &= F(x).\end{aligned}$$

Thus, $\mu = \lambda \circ T^{-1}$ has distribution function F.

This proves Theorem 14 for $d = 1$. It also proves Theorem 13 for $d = 1$, since, if we started with a measure μ and $F = F_\mu$, then we produced the map T under which Lebesgue measure pushes forward to μ.

11. Higher dimensions

The following (sketch of) proof of Theorem 14 applies to any dimension.

Proof of Theorem 14. We already showed uniqueness.

To show the existence, we may repeat the Caratheodary construction. We just sketch the starting point. Let $\mathcal{S}_d := \{I_1 \times \ldots \times I_d : I_j \in \mathcal{S}_1\}$, where $\mathcal{S}_1$ is the collection of left-open, right-closed

[1]The fact is nevertheless true. Any Borel probability measure on a complete and separable metric space (eg., $\mathbb{R}^d$) is the push-forward of Lebesgue measure on $[0, 1]$ under some measurable function. For $\mathbb{R}^d$, this can be shown using the one-dimensional result, but it will need us to develop the notion of conditional probability.

intervals in $\mathbb{R}$ (including those of the form $(-\infty, a]$ and (a, ∞)). Then $\mathcal{S}_d$ is a π-system and the algebra generated by it can be described explicitly as

$$\mathcal{A}_d := \left\{ \bigsqcup_{k=1}^{n} A_k : n \geq 0,\ A_k \in \mathcal{S}_d \text{ are pairwise disjoint} \right\}.$$

Given $F : \mathbb{R}^d \to [0,1]$ as in the statement of the theorem, we define $\mu : \mathcal{A}_d \to [0,1]$ as follows. First define it on $\mathcal{S}_d$ by setting

$$\mu\left((a_1, b_1] \times \ldots \times (a_d, b_d]\right) = \sum_{\substack{c_i \in \{a_i, b_i\} \\ 1 \leq i \leq d}} \pm F(c_1, \ldots, c_d)$$

where the signs must be appropriately chosen. For example, in $d = 1$, we set $\mu(a, b] = F(b) - F(a)$ while in $d = 2$, we set $\mu((a_1, b_1] \times (a_2, b_2]) = F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2)$. In general, the sign must be negative if there are an odd number of j for which $c_j \neq b_j$.

Then, for $A \in \mathcal{A}_d$, write $A = A_1 \sqcup \ldots \sqcup A_n$ with $A_i \in \mathcal{S}_d$ and define $\mu(A) = \mu(A_1) + \ldots + \mu(A_n)$.

The main part of the proof (which we skip) would be to check that μ is countably additive on the algebra $\mathcal{A}_d$ (it takes a bit of work). Then, invoke the result of Caratheodary to extend μ to $\mathcal{B}(\mathbb{R}^d)$ as a probability measure. By construction, the distribution function of μ will be F. ■

Next we turn to the proof of Theorem 13. To clarify the main idea, let us indicate how Lebesgue measure on $(0,1)^2$ can be got from Lebesgue measure on $(0,1)$.

Given $x \in (0,1)$, let $x = 0.t_1 t_2 t_3 \ldots$ be its binary expansion. Then define $y = 0.t_1 t_3 \ldots$ and $z = 0.t_2 t_4 \ldots$. Thus we get a mapping $x \mapsto (y, z)$ which goes from $(0,1)$ to $(0,1)^2$. It is not hard to see that this mapping is Borel measurable and the push-forward of Lebesgue measure on $(0,1)$ is the Lebesgue measure on $(0,1)^2$.

Convention: There are a couple of issues. Binary expansion is not uniquely defined. For example, $0.01011111\ldots$ and $0.01100000\ldots$ represent the same number. To avoid ambiguities, let us always take the expansion that has infinitely many ones. Then, for each $n \in \mathbb{Z}$, let $B_n : \mathbb{R} \to \{0,1\}$ be the function such that $B_n(x)$ is the nth digit in the binary expansion so that $x = \sum_{n \in \mathbb{Z}} B_n(x) 2^{-n}$. For any x, clearly $B_n(x) = 0$ if n is sufficiently negative, and our convention says that there are infinitely many $n \geq 1$ for which $B_n(x) = 1$.

Observe that each B_n is a step-function (where are the jumps and is it left or right continuous at those points?) and hence Borel measurable.

Proof of Theorem 13. For simplicity of notation, let $d = 2$ (write for yourself the case of general d). Define $T : \mathbb{R}^2 \to \mathbb{R}$ as

$$T(x, y) = \sum_{n \in \mathbb{Z}} \frac{B_n(x)}{2^{2n-1}} + \frac{B_n(y)}{2^{2n}}.$$

In words, $T(x,y)$ is got by interlacing the binary expansions of x and y. Clearly the sums are convergent and hence T is well-defined and Borel measurable (as it is a limit of finite sums of Borel measurable functions). Clearly T is injective, since we can recover x and y from the binary expansion of $T(x,y)$. Let $A \subseteq \mathbb{R}$ be the range of T so that $T : \mathbb{R}^2 \mapsto A$ is bijective.

We claim that A is a Borel set. To see this, first observe that

$$A^c = \{t \in \mathbb{R} : B_{2n}(t) = 1 \text{ for finitely many } n\} \bigcup \{t \in \mathbb{R} : B_{2n-1}(t) = 1 \text{ for finitely many } n\}.$$

For any finite subset $F \subseteq \mathbb{Z}$, let

$$B_F = \{t : B_{2n}(t) = 0 \text{ for } n \notin F \text{ and } B_{2n}(t) = 1 \text{ for } n \in F\},$$

$$C_F = \{t : B_{2n-1}(t) = 0 \text{ for } n \notin F \text{ and } B_{2n}(t) = 1 \text{ for } n \in F\},$$

so that $A^c = \bigcup_F B_F \cup C_F$, a countable union. Thus, it suffices to show that B_F and C_F are Borel sets for each F. That is obvious since

$$B_F = \bigcap_{n \in F} B_{2n}^{-1}\{1\} \bigcap_{n \in \mathbb{Z}\setminus F} B_{2n}^{-1}\{0\},$$

$$C_F = \bigcap_{n \in F} B_{2n-1}^{-1}\{1\} \bigcap_{n \in \mathbb{Z}\setminus F} B_{2n-1}^{-1}\{0\},$$

and each B_n is Borel measurable. This proves the claim that A is a Borel set.

Lastly if we define $S : \mathbb{R} \to \mathbb{R}^2$ by $S(z) = (x,y)$ where $x = \sum_{n=1}^{\infty} \frac{B_{2n-1}(z)}{2^n}$ and $y = \sum_{n=1}^{\infty} \frac{B_{2n}(z)}{2^n}$, then it is clear that S is Borel measurable. Further, $S \Big|_A$ is precisely T^{-1}. Since A is Borel, this shows that for any $C \in \mathcal{B}(\mathbb{R}^2)$, we get that $(T^{-1})^{-1}(C) = S^{-1}(C) \cap A$ is also a Borel set. Hence T^{-1} is Borel measurable.

Thus $T : \mathbb{R}^2 \to A$ is a bijection and both T and T^{-1} are Borel-measurable. Hence, give a probability measure μ on $\mathbb{R}^2$, the push-forward $\nu = \mu \circ T^{-1}$ is a Borel measure on $\mathbb{R}$. We know that $\nu = \lambda \circ h^{-1}$ for some Borel measurable $h : (0,1) \to \mathbb{R}$. Thus, $\mu = \lambda \circ h^{-1} \circ T$ or in words, the map $h \circ T^{-1} : (0,1) \to \mathbb{R}^2$ pushes the Lebesgue measure forward to the given measure μ. ■

12. Examples of probability measures in Euclidean space

There are many important probability measures that occur frequently in probability and in the real world. We give some examples below and expect you to familiarize yourself with each of them.

Example 16

The examples below have CDFs of the form $F(x) = \int_{-\infty}^{x} f(t)dt$ where f is a non-negative integrable function with $\int f = 1$. In such cases f is called the *density* or pdf (probability density function). Clearly F is continuous and non-decreasing and tends to 0 and 1 at $+\infty$ and

$-\infty$ respectively. Hence, there do exist probability measures on $\mathbb{R}$ with the corresponding density.

(1) *Normal distribution*. For fixed $a \in \mathbb{R}$ and $\sigma^2 > 0$, $N(a, \sigma^2)$ is the probability measure on $\mathbb{R}$ with density $\frac{1}{\sigma\sqrt{2\pi}} e^{-(x-a)^2/2\sigma^2} du$. F is clearly increasing and continuous and $F(-\infty) = 0$. That $F(+\infty) = 1$ is not so obvious but true!

(2) *Gamma distribution* with shape parameter $\alpha > -1$ and scale parameter $\lambda > 0$ is the probability measure with density $f(x) = \frac{1}{\Gamma(\alpha)} \lambda^\alpha x^{\alpha-1} e^{-\lambda x}$ for $x > 0$.

(3) *Exponential distribution*. Exponential(λ) is the probability measure with density $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ and $f(x) = 0$ if $x < 0$. This is a special case of Gamma distribution, but important enough to have its own name.

(4) *Beta distribution*. For parameters $a > -1$, $b > -1$, the Beta(a, b) distribution is the probability measure with density $B(a, b)^{-1} x^{a-1} (1 - x)^{b-1}$ for $x \in [0, 1]$. Here $B(a, b)$ is the beta function, defined as the constant that makes the integral to be 1. It can be shown to be equal to $\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

(5) *Uniform distribution* on $[a, b]$ is the probability measure with density $f(x) = \frac{1}{b-a}$ for $x \in [a, b]$. For example, with $a = 0, b = 1$, this is a special case of the Beta distribution.

(6) *Cauchy distribution*. This is the probability measure with density $\frac{1}{\pi(1+x^2)}$ on the whole line. Unlike all the previous examples, this distribution has "heavy tails"

You may have seen the following discrete probability measures. They are very important too and will recur often.

Example 17

The examples below have CDFs of the form $F(x) = \sum_{u_i \leq x} p(x_i) dt$, where $\{x_i\}$ is a fixed countable set, and $p(x_i)$ are non-negative numbers that add to one. In such cases p is called the pmf (probability mass function). and from what we have shown, there do exist probability measures on $\mathbb{R}$ with the corresponding density or CDF.

(1) *Binomial distribution*. Binomial(n, p), with $n \in \mathbb{N}$ and $p \in [0, 1]$, has the pmf $p(k) = \binom{n}{k} p^k q^{n-k}$ for $k = 0, 1, \ldots, n$.

(2) *Bernoulli distribution*. $p(1) = p$ and $p(0) = 1 - p$ for some $p \in [0, 1]$. Same as Binomail$(1, p)$.

(3) *Poisson(λ)* distribution with parameter $\lambda \geq 0$ has probability measuref $p(k) = e^{-\lambda}\frac{\lambda^k}{k!}$ for $k = 0, 1, 2, \ldots$.

(4) *Geometric(p)* distribution with parameter $p \in [0, 1]$ has probability measuref $p(k) = q^k p$ for $k = 0, 1, 2, \ldots$.

All the measures we mentioned so far are in one dimension. Among multi-variate ones, we mention one important example.

Example 18: Multivariate normal distribution

Let $\mu \in \mathbb{R}^d$ and Σ be a $d \times d$ symmetric, positive-definite matrix. Then,

$$f(x) := \frac{1}{(2\pi)^{d/2}\sqrt{\det(\Sigma)}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^t\Sigma^{-1}(\mathbf{x}-\mu)\right\}$$

is a probability density on $\mathbb{R}^d$. The probability measure with distribution function given by

$$F(x_1, \ldots, x_d) := \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_d} f(t_1, \ldots, t_d) dt_1 \ldots dt_d$$

is called the multi-variate normal distribution with mean vector μ and covariance matrix Σ (we are yet to define what mean and covariance means, but once defined this terminology will be justified).

Exercise 12

In each of the above examples, try to find a transformation $T : (0, 1) \to \mathbb{R}$ that pushes Lebesgue measure forward to the given probability measure. Implement this on a computer to generate random numbers from these distributions using a random number generator that outputs uniform random numbers in $[0, 1]$.

13. A METRIC ON THE SPACE OF PROBABILITY MEASURES ON $\mathbb{R}^d$

What kind of space is $\mathcal{P}(\mathbb{R}^d)$, the space of Borel on $\mathbb{R}^d$? It is clearly a convex set (this is true for the space of probability measures on any measurable space). We want to measure closeness of two probability distributions. Two possible definitions come to mind.

(1) For $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, define $D_1(\mu, \nu) := \sup_{A \in \mathcal{B}_d} |\mu(A) - \nu(A)|$. Since μ and ν are functions on the Borel σ-algebra, this is just their supremum distance, usually called the *total variation distance*. It is easy to see that D_1 is indeed a metric on $\mathcal{P}(\mathbb{R}^d)$.

One shortcoming of this metric is that if μ is a discrete measure and ν is a measure with density, then $D_1(\mu, \nu) = 1$. But we shall be interested in talking about discrete measures

approximating continuous ones (as in central limits theorem, if you have heard of it). The metric D_1 is too strong for this purpose.

(2) We can restrict the class of sets over which we take the supremum. For instance, taking all semi-infinite intervals, we define the *Kolmogorov-Smirnov* distance

$$D_2(\mu,\nu) = \sup_{x\in\mathbb{R}^d} |F_\mu(x) - F_\nu(x)|.$$

If two CDFs are equal, the corresponding measures are equal. Hence D_2 is also a genuine metric on $\mathcal{P}(\mathbb{R}^d)$.

Clearly $D_2(\mu,\nu) \leq D_1(\mu,\nu)$, hence D_2 is weaker than D_1. Unlike with D_1, it is possible to have discrete measures converging in D_2 to a continuous one, see Exercise 13. But it is still too strong.

For example, if $a \neq b$ are points in $\mathbb{R}^n$, then it is easy to see that $D_1(\delta_a,\delta_b) = D_2(\delta_a,\delta_b) = 1$. Thus, even when $a_n \to a$ in $\mathbb{R}^d$, we do not get convergence of δ_{a_n} to δ_a in these metrics. This is an undesirable feature (why? Let us just say that we would like the embedding $\mathbb{R} \mapsto \mathcal{P}(\mathbb{R})$ defined by $a \mapsto \delta_a$ to be continuous).

Thus, we would like a weaker metric, where more sequences converge. The problem with the earlier two definitions is that they compare $\mu(A)$ with $\nu(A)$. The next definition allows the set to change a little.

Definition 7

For $\mu,\nu \in \mathcal{P}(\mathbb{R}^d)$, define the *Lévy distance* between them as (here $\mathbf{1} = (1,1,\ldots,1)$)

$$d(\mu,\nu) := \inf\{u > 0 : F_\mu(x+u\mathbf{1}) + u \geq F_\nu(x),\ F_\nu(x+u\mathbf{1}) + u \geq F_\mu(x)\ \forall x \in \mathbb{R}^d\}.$$

If $d(\mu_n,\mu) \to 0$, we say that μ_n converges in distribution or converges weakly to μ and write $\mu_n \xrightarrow{d} \mu$. [...breathe slowly and meditate on this definition for a few minutes...]

Remark 7

Although we shall not use it, we mention how a distance is defined on $\mathcal{P}(X)$ for a metric space X (it is called *Lévy-Prohorov distance*). For $\mu,\nu \in \mathcal{P}(X)$

$$d(\mu,\nu) := \inf\{t > 0 : \mu(A^{(t)}) + t \geq \nu(A) \text{ and } \nu(A^{(t)}) + t \geq \mu(A) \text{ for all closed } A \subseteq X\}.$$

Here $A^{(t)}$ is the set of all points in X that are within distance t of A. This makes it clear that we do not directly compare the measures of a given set, but if $d(\mu,\nu) < t$, it means that whenever μ gives a certain measure to a set, then ν should give nearly that much (nearly means, allow t amount less) measure to a t-neighbourhood of A.

As an example, if $a, b \in \mathbb{R}^d$, then check that $d(\delta_a, \delta_b) \leq (\max_i |b_i - a_i|) \wedge 1$. Hence, if $a_n \to a$, then $d(\delta_{a_n}, \delta_a) \to 0$. Recall that δ_{a_n} does not converge to δ_a in D_1 or D_2.

Exercise 13

Let $\mu_n = \frac{1}{n}\sum_{k=1}^{n} \delta_{k/n}$. Show directly by definition that $d(\mu_n, \lambda) \to 0$. Show also that $D_2(\mu_n, \lambda) \to 0$ but $D_1(\mu_n, \lambda)$ does not go to 0.

The definition is rather unwieldy in checking convergence. The following proposition gives the criterion for convergence in distribution in terms of distribution functions.

Proposition 15

$\mu_n \xrightarrow{d} \mu$ if and only if $F_{\mu_n}(x) \to F_\mu(x)$ for all continuity points x of F_μ.

Proof. Suppose $\mu_n \xrightarrow{d} \mu$. Let $x \in \mathbb{R}^d$ and fix $u > 0$. Then for large enough n, we have $F_\mu(x + u\mathbf{1}) + u \geq F_{\mu_n}(x)$, hence $\limsup F_{\mu_n}(x) \leq F_\mu(x + u\mathbf{1}) + u$ for all $u > 0$. By right continuity of F_μ, we get $\limsup F_{\mu_n}(x) \leq F_\mu(x)$. Further, $F_{\mu_n}(x) + u \geq F_\mu(x - u\mathbf{1})$ for large n, hence $\liminf F_{\mu_n}(x) \geq F_\mu(x - u)$ for all u. If x is a continuity point of F_μ, we can let $u \to 0$ and get $\liminf F_{\mu_n}(x) \geq F_\mu(x)$. Thus $F_{\mu_n}(x) \to F_\mu(x)$.

For the converse, for simplicity let $d = 1$. Suppose $F_n \to F$ at all continuity points of F. Fix any $u > 0$. Find $x_1 < x_2 < \ldots < x_m$, continuity points of F, such that $x_{i+1} \leq x_i + u$ and such that $F(x_1) < u$ and $1 - F(x_m) < u$. This can be done because continuity points are dense. Now use the hypothesis to fix N so that $|F_n(x_i) - F(x_i)| < u$ for each $i \leq m$ and for $n \geq N$. Henceforth, let $n \geq N$.

If $x \in \mathbb{R}$, then either $x \in [x_{j-1}, x_j]$ for some j or else $x < x_1$ or $x > x_1$. First suppose $x \in [x_{j-1}, x_j]$. Then

$$F(x + u) \geq F(x_j) \geq F_n(x_j) - u \geq F_n(x) - u, \qquad F_n(x + u) \geq F_n(x_j) \geq F(x_j) - u \geq F(x) - u.$$

If $x < x_1$, then $F(x + u) + u \geq u \geq F(x_1) \geq F_n(x_1) - u$. Similarly the other requisite inequalities, and we finally have

$$F_n(x + 2u) + 2u \geq F(x) \text{ and } F(x + 2u) + 2u \geq F_n(x).$$

Thus $d(\mu_n, \mu) \leq 2u$. Hence $d(\mu_n, \mu) \to 0$. ∎

Example 19

Again, let $a_n \to a$ in $\mathbb{R}$. Then $F_{\delta_{a_n}}(t) = 1$ if $t \geq a_n$ and 0 otherwise while $F_{\delta_a}(t) = 1$ if $t \geq a$ and 0 otherwise. Thus, $F_{\delta_{a_n}}(t) \to F_{\delta_a}(t)$ for all $t \neq a$ (just consider the two cases $t < a$ and

$t > a$). This example also shows the need for excluding discontinuity points of the limiting distribution function. Indeed, $F_{\delta_{a_n}}(a) = 0$ (if $a_n \neq a$) but $F_{\delta_a}(a) = 1$.

Observe how much easier it is to check the condition in the theorem rather than the original definition! Many books use the convergence at all continuity points of the limit CDF as the definition of convergence in distribution. But we defined it via the Lévy metric because we are familiar with convergence in metric spaces and this definition shows that convergence in distribution in not anything more exotic (as it might sound from the other definition).

Exercise 14

If $a_n \to 0$ and $b_n^2 \to 1$, show that $N(a_n, b_n^2) \xrightarrow{d} N(0,1)$ (recall that $N(a, b^2)$ is the Normal distribution with parameters $a \in \mathbb{R}$ and $b^2 > 0$).

Question: In class, Milind Hegde raised the following question. If we define (write in one dimension for notational simplicity)

$$d'(\mu, \nu) = \inf\{t > 0 : F_\mu(x+t) \geq F_\nu(x) \text{ and } F_\nu(x+t) \geq F_\mu(x) \text{ for all } x\},$$

how different is the resulting metric from the Lévy metric? In other words, is it necessary to allow an extra additive t to $F_\mu(x+t)$?

It does make a difference! Suppose μ, ν are two probability measures on $\mathbb{R}$ such that $\mu(K_0) = 1$ for some compact set K_0 and $\nu(K) < 1$ for all compact sets K. Then, if x is large enough so that $x > y$ for all $y \in K_0$, then $F_\nu(x+t) < 1 = F_\mu(x)$ for any $t > 0$. Hence, $d'(\mu, \nu) > t$ for any t implying that $d'(\mu, \nu) = \infty$.

Now, it is not a serious problem if a metric takes the value ∞. We can replace d' by $d''(\mu, \nu) = d'(\mu, \nu) \wedge 1$ or $d'''(\mu, \nu) = d(\mu, \nu)/(1 + d(\mu, \nu))$ which gives metrics that are finite everywhere but are such that convergent sequences are the same as in d' (i.e., $d'(\mu_n, \mu) \to 0$ if and only if $d''(\mu_n, \mu) \to 0$).

But the issue is that measures with compact support can never converge to a measure without compact support. For example, if X has exponential distribution and $X_k = X \wedge k$, then the distribution of X_k does not converge to the distribution of X in the metric d'. However, it is indeed the case that the convergence happens in the metric d. Thus the two metrics are not equivalent [2].

[2] In class I wrongly claimed that for probability measures on a compact set in place of the whole real line, eg., $\mathcal{P}([-1,1])$, convergence in d' and in d are equivalent. Chirag Igoor showed me the following counter-example. Let

14. Compact subsets in the space of probability measure on Euclidean spaces

Often we face problems like the following. A functional $L : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ is given, and we would like to find the probability measure μ that minimizes $L(\mu)$. By definition, we can find nearly optimal probability measures μ_n satisfying $L(\mu_n) - \frac{1}{n} \leq \inf_\nu L(\nu)$. Then we might expect that *if* the sequence μ_n (or a subsequence of it) converged to a probability measure μ, then μ might be the optimal solution we are searching for. This motivates us to characterize compact subsets of $\mathcal{P}(\mathbb{R}^d)$, so that existence of convergent subsequences can be asserted.

Looking for a convergent subsequence: Let μ_n be a sequence in $\mathcal{P}(\mathbb{R}^d)$. We would like to see if a convergent subsequence can be extracted. Towards this direction, we prove the following lemma. We emphasize the idea of proof (a diagonal argument) which recurs in many contexts.

Lemma 16

[Helly's selection principle] Let F_n be a sequence distribution functions on $\mathbb{R}^d$. Then, there exists a subsequence $\{n_\ell\}$ and a non-decreasing, right continuous functon $F : \mathbb{R}^d \to [0,1]$ such that $F_{n_\ell}(x) \to F(x)$ if x is a continuity point of F.

Proof. Fix a dense subset $S = \{x_1, x_2, \ldots\}$ of $\mathbb{R}^d$. Then, $\{F_n(x_1)\}$ is a sequence in $[0,1]$. Hence, we can find a subsequence $\{n_{1,k}\}_k$ such that $F_{n_{1,k}}(x_1)$ converges to some number $\alpha_1 \in [0,1]$. Then, extract a further subsequence $\{n_{2,k}\}_k \subseteq \{n_{1,k}\}_k$ such that $F_{n_{2,k}}(x_2) \to \alpha_2$, another number in $[0,1]$. Of course, we also have $F_{n_{2,k}}(x_1) \to \alpha_1$. Continuing this way, we get numbers $\alpha_j \in [0,1]$ and subsequences $\{n_{1,k}\} \supset \{n_{2,k}\} \supset \ldots \{n_{\ell,k}\} \ldots$ such that for each ℓ, as $k \to \infty$, we have $F_{n_{\ell,k}}(x_j) \to \alpha_j$ for each $j \leq \ell$.

The *diagonal subsequence* $\{n_{\ell,\ell}\}$ is ultimately the subsequence of each of the above obtained subsequences and therefore, $F_{n_{\ell,\ell}}(x_j) \to \alpha_j$ as $\ell \to \infty$, for each j. Henceforth, write n_ℓ instead of $n_{\ell,\ell}$.

To get a function on the whole line, set $F(x) := \inf\{\alpha_j : j \text{ for which } x_j > x\}$. F is well defined, takes values in $[0,1]$ and is non-decreasing. It is also right-continuous, because if $y_n \downarrow y$, then for any j for which $x_j > y$, it is also true that $x_j > y_n$ for sufficiently large n. Thus $\lim_{n\to\infty} F(y_n) \leq \alpha_j$.

$\mu = \delta_1$ and for each n define

$$F_n(x) = \begin{cases} 0 & \text{if } x < 0, \\ x/n & \text{if } 0 \leq x < 1, \\ 1 & \text{if } x \geq 1. \end{cases}$$

Then, $F_n(x) \to F_\mu(x)$ for each x and hence the corresponding measures converge to μ in Lévy metric. But the convergence fails in d'. To see this, take any $x > 0$ and observe that if $F_\mu(0.5 + t) \geq F_{\mu_n}(0.5)$, then we must have $t \geq 0.5$. As this is true for every n, it follows that μ_n does not converge to μ in d'.

Take infimum over all j such that $x_j > y$ to get $\lim_{n\to\infty} F(y_n) \leq F(y)$. Of course $F(y) \leq \lim F(y_n)$ as F is non-decreasing. This shows that $\lim F(y_n) = F(y)$ and hence F is right continuous.

Lastly, we claim that if y is any continuity point of F, then $F_{n_\ell}(y) \to F(y)$ as $\ell \to \infty$. To see this, fix $\delta > 0$. Find i, j such that $y - \delta < x_i < y < x_j < y + \delta$. Therefore

$$\liminf F_{n_\ell}(y) \geq \lim F_{n_\ell}(x_i) = \alpha_i \geq F(y - \delta)$$

$$\limsup F_{n_\ell}(y) \leq \lim F_{n_\ell}(x_j) = \alpha_j \leq F(y + \delta).$$

In each line, the first inequalities are by the increasing nature of CDFs, and the second inequalities are by the definition of F. Thus

$$F(y-) \leq \liminf F_{n_\ell}(y) \leq \limsup F_{n_\ell}(y) \leq F(y)$$

for all $y \in \mathbb{R}$. If $F(y-) = F(y)$, then it follows that $\lim F_{n_\ell}(y)$ exists and equals $F(y)$. ■

The Lemma does not say that F is a CDF, because in general it is not!

Example 20

Consider δ_n. Clearly $F_{\delta_n}(x) \to 0$ for all x if $n \to +\infty$ and $F_{\delta_n}(x) \to 1$ for all x if $n \to -\infty$. Even if we pass to subsequences, the limiting function is identically zero or identically one, and neither of these is a CDF of a probability measure The problem is that mass escapes to infinity. To get weak convergence to a probability measure, we need to impose a condition to avoid this sort of situation.

Definition 8

A family of probability measure $\mathcal{A} \subseteq \mathcal{P}(\mathbb{R}^d)$ is said to be *tight* if for any $\epsilon > 0$, there is a compact set $K_\epsilon \subseteq \mathbb{R}^d$ such that $\mu(K_\epsilon) \geq 1 - \epsilon$ for all $\mu \in \mathcal{A}$.

Example 21

Suppose the family has only one probability measure μ. Since $[-n, n]^d$ increase to $\mathbb{R}^d$, given $\epsilon > 0$, for a large enough n, we have $\mu([-n, n]^d) \geq 1 - \epsilon$. Hence $\{\mu\}$ is tight. If the family is finite, tightness is again clear.

Take $d = 1$ and let μ_n be probability measures with $F_n(x) = F(x - n)$ (where F is a fixed CDF), then $\{\mu_n\}$ is not tight. This is because given any $[-M, M]$, if n is large enough, $\mu_n([-M, M])$ can be made arbitrarily small. Similarly $\{\delta_n\}$ is not tight.

We now characterize compact subsets of $\mathcal{P}(R^d)$ in the following theorem. As $\mathcal{P}(\mathbb{R}^d)$ is a metric space, compactness is equivalent to sequential compactness and we phrase the theorem in terms of sequential compactness.

Theorem 17

Let $\mathcal{A} \subseteq \mathcal{P}(\mathbb{R}^d)$. Then, the following are equivalent.

(1) Every sequence in $\mathcal{A}$ has a convergent subsequence in $\mathcal{P}(\mathbb{R}^d)$.

(2) $\mathcal{A}$ is tight.

Proof. Let us take $d = 1$ for simplicity of notation.

(1) Assume that $\mathcal{A}$ is tight. Then any sequence $(\mu_n)_n$ in $\mathcal{A}$ is also tight. By Lemma 16, there is a subsequence $\{n_\ell\}$ and a non-decreasing right continuous function F (taking values in $[0,1]$) such that $F_{n_\ell}(x) \to F(x)$ for all continuity points x of F.

Fix $A > 0$ such that $\mu_{n_\ell}[-A, A] \geq 1 - \epsilon$ and such that A is a continuity point of F. Then, $F_{n_\ell}(-A) \leq \epsilon$ and $F_{n_\ell}(A) \geq 1 - \epsilon$ for every n and by taking limits we see that $F(-A) \leq \epsilon$ and $F(A) \geq 1 - \epsilon$. Thus $F(+\infty) = 1$ and $F(-\infty) = 0$. This shows that F is a CDF and hence $F = F_\mu$ for some $\mu \in \mathcal{P}(\mathbb{R}^d)$. By Proposition 15 it also follows that $\mu_{n_\ell} \xrightarrow{d} \mu$.

(2) Assume that $\mathcal{A}$ is not tight. Then, there exists $\epsilon > 0$ such that for any k, there is some $\mu_k \in \mathcal{A}$ such that $\mu_k([-k, k]) < 1 - 2\epsilon$. In particular, either $F_{\mu_k}(k) \leq 1 - \epsilon$ or/and $F_{\mu_k}(-k) \geq \epsilon$. We claim that no subsequence of $(\mu_k)_k$ can have a convergent subsequence.

To avoid complicating the notation, let us show that the whole sequence does not converge and leave you to rewrite the same for any subsequence. There are infinitely many k for which $F_{\mu_k}(-k) \geq \epsilon$ or there are infinitely many k for which $F_{\mu_k}(k) \geq 1 - \epsilon$. Suppose the former is true. Then, for any $x \in \mathbb{R}$, since $-k < x$ for large enough k, we see that $F_{\mu_k}(x) \geq F_{\mu_k}(-k) \geq \epsilon$ for large enough k. This means that if F_{μ_k} converge to some F (at continuity points of F), then $F(x) \geq \epsilon$ for all x. Thus, F cannot be a CDF and hence μ_k does not have a limit. ■

Exercise 15

Adapt this proof for $d \geq 2$.

15. Expectation

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. We define *Expectation* or *Lebesgue integral* of real-valued random variables in three steps.

(1) If X can be written as $X = \sum_{i=1}^n c_i \mathbf{1}_{A_i}$ for some $A_i \in \mathcal{F}$, we say that X is a *simple r.v.*. We define its *expectation* to be $\mathbf{E}[X] := \sum_{i=1}^n c_i \mathbf{P}(A_i)$.

(2) If $X \geq 0$ is a random variable, we define $\mathbf{E}[X] := \sup\{\mathbf{E}[S] : 0 \leq S \leq X, \ \ S \text{ is a simple r.v.}\}$, which is either a non-negative number or $+\infty$.

(3) If X is any real-valued random variable, let $X_+ := X\mathbf{1}_{X\geq 0}$ and $X_- := -X\mathbf{1}_{X<0}$ so that $X = X_+ - X_-$ (also observe that $X_+ + X_- = |X|$). If both $\mathbf{E}[X_+]$ and $\mathbf{E}[X_-]$ are finite, we say that X is integrable (or that its expectation exists) and define $\mathbf{E}[X] := \mathbf{E}[X_+] - \mathbf{E}[X_-]$.

Naturally, there are some arguments needed to complete these steps. We elaborate a little. But full details are left to measure theory class (or consult any measure theory book, eg., Dudley's *Real analysis and probability*).

(1) In the first step, one should check that $\mathbf{E}[X]$ is well-defined. This is necessary because a simple random variable can be represented as $\sum_{i=1}^n c_i\mathbf{1}_{A_i}$ in many ways. Finite additivity of $\mathbf{P}$ is used to show this. It helps to note that there is a unique way to write X in this form so that the sets A_k are pairwise disjoint and numbers c_k are distinct.

(2) In addition, check that the expectation operator defined on simple random variables has the following properties.
   (a) *Linearity:* If X, Y are simple random variables, then $\mathbf{E}[\alpha X + \beta Y] = \alpha\mathbf{E}[X] + \beta\mathbf{E}[Y]$ for all $\alpha, \beta \in \mathbb{R}$.
   (b) *Positivity:* If $X \geq 0$ (this means that $X(\omega) \geq 0$ for all $\omega \in \Omega$), then $\mathbf{E}[X] \geq 0$.

(3) Then go to the second step and define expectation of non-negative random variables. Again we must check that linearity and positivity are preserved. It is clear that $\mathbf{E}[\alpha X] = \alpha\mathbf{E}[X]$ if $X \geq 0$ is a r.v and α is a non-negative real number (why?). One can also easily see that $\mathbf{E}[X+Y] \geq \mathbf{E}[X] + \mathbf{E}[Y]$ using the definition. To show that $\mathbf{E}[X+Y] = \mathbf{E}[X] + \mathbf{E}[Y]$, it is necessary to use countable additivity of $\mathbf{P}$ in the following form.

Theorem 18: Monotone convergence theorem - provisional version

If S_n are non-negative simple r.v.s that increase to X (i.e., $S_n(\omega) \uparrow X(\omega)$ for each $\omega \in \Omega$), then $\mathbf{E}[S_n]$ increases to $\mathbf{E}[X]$.

If $S_n \uparrow X$ and $T_n \uparrow Y$, then $S_n + T_n \uparrow X + Y$ (and $S_n + T_n$ is simple if S_n and T_n are), hence we get the conclusion that $\mathbf{E}[X+Y] = \mathbf{E}[X] + \mathbf{E}[Y]$ for non-negative random variables. To avoid vacuous statements, we note that there do exist simple random variables S_n, T_n that increase to X, Y. For example, we can take

$$S_n(\omega) = \sum_{k=0}^{2^{2n}} \frac{k}{2^n}\mathbf{1}_{X(\omega)\in[k2^{-n},(k+1)2^{-n})}.$$

(4) It is convenient to allow a non-negative random variable to take the value $+\infty$ but adopt the convention that $0 \cdot \infty = 0$. That is, infinite value on a set of zero probability does not

matter in computing expectations. Of course, if a non-negative random variable takes the value $+\infty$ on set of positive probability, then $\mathbf{E}[X] = +\infty$ (follows from the definition).

(5) In step 3, one assumes that both $\mathbf{E}[X_+]$ and $\mathbf{E}[X_-]$ are finite, which is equivalent to assuming that $\mathbf{E}[|X|] < \infty$ (because $|X| = X_+ + X_-$). Such random variables are said to be *integrable* or *absolutely integrable*. For an integrable random variable X, we define $\mathbf{E}[X] := \mathbf{E}[X_+] - \mathbf{E}[X_-]$.

(6) Finally argue that on the collection of all integrable random variables on the given probability space, the expectation operator is linear and positive.

Convention: Let us say "$X = Y$ $a.s$" or "$X < Y$ $a.s$" etc., to mean that $\mathbf{P}(X = Y) = 1$ or $\mathbf{P}(X < Y) = 1$ etc. We may also use a.e. (almost everywhere) or w.p.1 (with probability one) in place of a.s (almost surely). More generally, if we write $[...xyz...]$, $a.s.$, we mean that whatever event is describe in $[...xyz...]$ has probability equal to 1. For example, the statement

$$X_n \to X \quad a.s.$$

just means the same as the statement

$$\mathbf{P}\{\omega : \lim_{n\to\infty} X_n(\omega) \text{ exists and is equal to } X(\omega)\} = 1.$$

Just as we ignore events having zero probability, we also do not usually distinguish two random variables that are equal almost surely. For example, is $X = Y$, $a.s.$, then their distributions $\mathbf{P} \circ X^{-1}$ and $\mathbf{P} \circ Y^{-1}$ are the same (why?). Similarly, if X is integrable, then so is Y and $\mathbf{E}[Y] = \mathbf{E}[X]$. For all probability questions of interest, the two random variables give the same answer and so they are essentially the same.

Summary: Given a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, let $L^1(\Omega, \mathcal{F}, \mathbf{P})$ be the collection of all integrable random variables on Ω. Then, the expectation operator $\mathbf{E} : L^1(\Omega, \mathcal{F}, \mathbf{P}) \to \mathbb{R}$ has the following properties.

(1) *Linearity:* If X, Y are integrable, then for any $\alpha, \beta \in \mathbb{R}$, the random variable $\alpha X + \beta Y$ is also integrable and $\mathbf{E}[\alpha X + \beta Y] = \alpha \mathbf{E}[X] + \beta \mathbf{E}[Y]$.

(2) *Positivity:* $X \geq 0$ implies $\mathbf{E}[X] \geq 0$. Further, if $X \geq 0$ and $\mathbf{P}(X = 0) < 1$, then $\mathbf{E}[X] > 0$.

A useful corollary of positivity is that whenever $X \leq Y$ and $\mathbf{E}[X], \mathbf{E}[Y]$ exist, then $\mathbf{E}[X] \leq \mathbf{E}[Y]$ with equality if and only if $X = Y$ a.s.

(3) $|\mathbf{E}[X]| \leq \mathbf{E}[|X|]$.

(4) $\mathbf{E}[\mathbf{1}_A] = \mathbf{P}(A)$ for $A \in \mathcal{F}$. In particular, $\mathbf{E}[1] = 1$.

Theorem 19: Monotone convergence theorem (MCT)

Suppose X_n, X are non-negative r.v.s and $X_n \uparrow X$ a.s. Then $\mathbf{E}[X_n] \uparrow \mathbf{E}[X]$. (valid even when $\mathbf{E}[X] = +\infty$).

Theorem 20: Fatou's lemma

Let X_n be non-negative r.v.s. Then $\mathbf{E}[\liminf X_n] \leq \liminf \mathbf{E}[X_n]$.

Theorem 21: Dominated convergence theorem (DCT)

Let $|X_n| \leq Y$ where Y is a non-negative r.v. with $\mathbf{E}[Y] < \infty$. If $X_n \to X$ a.s., then, $\mathbf{E}[|X_n - X|] \to 0$ and hence we also get $\mathbf{E}[X_n] \to \mathbf{E}[X]$.

Assuming MCT, the other two follow easily. For example, to prove Fatou's lemma, just define $Y_n = \inf_{n \geq k} X_n$ and observe that Y_ks increase to $\liminf X_n$ a.s and hence by MCT $\mathbf{E}[Y_k] \to \mathbf{E}[\liminf X_n]$. Since $X_n \geq Y_n$ for each n, we get $\liminf \mathbf{E}[X_n] \geq \liminf \mathbf{E}[Y_n] = \mathbf{E}[\liminf X_n]$.

To prove DCT, first note that $|X_n| \leq Y$ and $|X| \leq Y$ a.s. Consider the sequence of non-negative r.v.s $2Y - |X_n - X|$ that converges to $2Y$ a.s. Then, apply Fatou's lemma to get

$$\mathbf{E}[2Y] = \mathbf{E}[\liminf(2Y - |X_n - X|)] \leq \liminf \ \mathbf{E}[2Y - |X_n - X|] = \mathbf{E}[2Y] - \limsup \mathbf{E}[|X_n - X|].$$

Thus $\limsup \mathbf{E}[|X_n - X|] = 0$. Further, $|\mathbf{E}[X_n] - \mathbf{E}[X]| \leq \mathbf{E}[|X_n - X|] \to 0$.

We omit the proof of MCT. But let us understand the conditions in these statements by giving examples that violate the conditions and for which the conclusions are false.

Example 22

Consider the probability space $([0,1], \mathcal{B}, \lambda)$. Let $f_n(t) = -\frac{1}{nt}$ and let $f(t) = 0$. Then, $f_n(t) \uparrow f(t)$ for all $t \neq 0$. However, $\mathbf{E}[f_n] = -\infty$ for each n and thus does not converge to $\mathbf{E}[f] = 0$. Thus, the conclusion of MCT is violated. But the conditions are not satisfied either, since f_n are not non-negative.

This is essentially the only way in which MCT can fail. Indeed, suppose that $X_n \uparrow X$ a.s. but X_n are not necessarily non-negative. Assume that $\mathbf{E}[(X_1)_-] < \infty$. Then, define $Y_n = X_n - X_1$ and $Y = X - X_1$. Clearly, $Y_n \geq 0$ $a.s.$ and $Y_n \uparrow Y$. Hence by MCT as stated above, $\mathbf{E}[Y_n] \uparrow \mathbf{E}[Y]$. But $\mathbf{E}[Y_n] = \mathbf{E}[X_n] - \mathbf{E}[X_1]$ and $\mathbf{E}[Y] = \mathbf{E}[X] - \mathbf{E}[X_1]$ (these statements are valid even if $\mathbf{E}[X_n]$ or $\mathbf{E}[X]$ is equal to ∞, since our assumption implies that $-\infty < \mathbf{E}[X_1] \leq +\infty$). Thus, MCT is valid even if we only assume that $X_n \uparrow X$ $a.s.$ and that $\mathbf{E}[(X_N)_-] < \infty$ for some N. In other words, for MCT to fail, we must have $\mathbf{E}[(X_n)_-] = +\infty$ for each n, as it happened in the above example.

The above example also shows how Fatou's lemma may be violated without the condition of $X_n \geq 0$ $a.s.$. We give another example, as unlike MCT, there are other ways in which Fatou's lemma may be violated.

Example 23

On the probability space $([0,1], \mathcal{B}, \lambda)$, define $f_n(t) = -n\mathbf{1}_{t \leq \frac{1}{n}}$ and $f(t) = 0$. Then $f_n \to f$ $a.s.$but $\mathbf{E}[f_n] = -1$ for all n while $\mathbf{E}[f] = 0$. If we reversed the signs, then $-f_n \geq 0$ and Fatou's lemma is indeed valid.

Clearly, Fatou's lemma implies that if $X_n \leq 0$, then $\mathbf{E}[\limsup X_n] \geq \limsup \mathbf{E}[X_n]$. A common mistake is to forget the reversed condition $X_n \leq 0$ which leads to wonderful conclusions like $0 > 1$. Lastly, an example where DCT fails.

Example 24

Again on the probability space $([0,1], \mathcal{B}, \lambda)$, define $f_n(t) = n\mathbf{1}_{t \leq \frac{1}{n}}$ and $f(t) = 0$. Then $f_n \to f$ $a.s.$, but $\mathbf{E}[f_n] = 1$ for all n but $\mathbf{E}[f] = 0$. DCT is not contradicted because there is no integrable random variable that dominates each f_n.
However, note that Fatou's lemma applies and is valid. Ideally we would like the conclusion of DCT (limit of expectations is equal to the expectation of the limit), but when that is not available, Fatou's may apply to give a one way inequality. You may see some similarity with the proof of Helly's theorem, where we show that a sequence of measures may lose some mass in the limit, but can never gain extra mass!

Here is a new way in which a random variable on a probability space gives rise to new probability measures on the same space.

Exercise 16

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and let $X \geq 0$ be a random variable with finite expectation. Define $\mathbf{Q} : \mathcal{F} \to \mathbb{R}_+$ by $\mathbf{Q}(A) = \frac{1}{\mathbf{E}[X]}\mathbf{E}[X\mathbf{1}_A]$. Show that $\mathbf{Q}$ is a probability measure on $\mathcal{F}$. Further, for any bounded random variable Y, we have $\mathbf{E_Q}[Y] = \frac{1}{\mathbf{E_P}[X]}\mathbf{E_P}[XY]$ (when we have more than one probability measure, we put a subscript to $\mathbf{E}$ to denote which measure we take expectations with respect to).

17. Lebesgue integral versus Riemann integral

Consider the probability space $([0,1], \bar{\mathcal{B}}, \lambda)$ (note that in this section we consider the Lebesgue σ-algebra, not Borel!) and a function $f : [0,1] \to \mathbb{R}$. Let

$$U_n := \frac{1}{2^n}\sum_{k=0}^{2^n-1} \max_{\frac{k}{2^n} \leq x \leq \frac{k+1}{2^n}} f(x), \qquad L_n := \frac{1}{2^n}\sum_{k=0}^{2^n-1} \min_{\frac{k}{2^n} \leq x \leq \frac{k+1}{2^n}} f(x)$$

be the upper and lower Riemann sums. Then, $L_n \leq U_n$ and U_n decrease with n while L_n increase. *If* it happens that $\lim U_n = \lim L_n$, we say that f is Riemann integrable and this common limit is defined to be the Riemann integral of f. The question of which functions are indeed Riemann integrable is answered precisely by[3]

Theorem 22: Lebesgue's theorem on Riemann integrals

A bounded function $f : [0,1] \to \mathbb{R}$ is Riemann integrable if and only if the set of discontinuity points of f has zero Lebesgue outer measure.

Next consider the Lebesgue integral $\mathbf{E}[f]$. For this we need f to be Lebesgue measurable in the first place. Clearly any bounded and measurable function is integrable (why?).

Further, we claim that if f is continuous a.e., then f is measurable. To see this, let $E \subseteq [0,1]$ be the set of discontinuity points of f. Then by assumption $\lambda_*(E) = 0$. Hence, E and all its subsets are Lebesgue measurable and have measure 0. Further, as E contains no interval, we can find a countable set $D \subseteq E^c$ that is dense in $[0,1]$. Let $A_s = \{x \in D : f(x) < s\}$, a countable set for any $s \in \mathbb{R}$ and hence measurable. Thus, for any $t \in \mathbb{R}$,

$$\{f \leq t\} = \{x \in E : f(x) \leq t\} \bigcup \left(\bigcap_{n \geq 1} (E^c \cap \overline{A}_{t+\frac{1}{n}}) \right).$$

This shows that $f < t$ is measurable.

Putting everything together, we see that Riemann integrable functions are also Lebesgue integrable. Further, if f is Riemann integrable, then its Riemann integral and Lebesgue integral agree. To see this, define

$$g_n(x) := \sum_{k=0}^{2^n-1} \left(\max_{\frac{k}{2^n} \leq x \leq \frac{k+1}{2^n}} f(x) \right) \mathbf{1}_{\frac{k}{2^n} \leq x \leq \frac{k+1}{2^n}}, \qquad h_n(x) := \sum_{k=0}^{2^n-1} \left(\min_{\frac{k}{2^n} \leq x \leq \frac{k+1}{2^n}} f(x) \right) \mathbf{1}_{\frac{k}{2^n} \leq x \leq \frac{k+1}{2^n}}$$

so that $\mathbf{E}[g_n] = U_n$ and $\mathbf{E}[h_n] = L_n$. Further, $g_n(x) \downarrow f(x)$ and $h_n(x) \uparrow f(x)$ at all continuity points of f. By MCT, $\mathbf{E}[g_n]$ and $\mathbf{E}[h_n]$ converge to $\mathbf{E}[f]$, while by the assumed Riemann integrability L_n and U_n converge to the Riemann integral of f. Thus the Lebesgue integral $\mathbf{E}[f]$ agrees with the Riemann integral.

In short, when a function is Riemann integrable, it is also Lebesgue integrable, and the integrals agree. But there are functions that are measurable but not a.e. continuous, for example, the function $\mathbf{1}_{\mathbb{Q} \cap [0,1]}$. Thus, Lebesgue integral is more powerful than Riemann integral. Henceforth in life, we shall always use the Lebesgue integral.

A natural question is whether there is an even more general way of defining an "integral"? To answer that, we need to know what we require out of an integral. Let us stick to function on $[0,1]$

[3] See Theorem 11.33 in Rudin's *Principles of mathematical analysis.*

for definiteness. Then we certainly want continuous functions to be integrable and the integral to satisfy linearity and positivity. Then, we have the following theorem of F. Riesz.

Theorem 23

Suppose $I : C[0,1] \to \mathbb{R}$ is a positive linear functional and $I(1) = 1$. That is, (1) $I(af + bg) = aI(f) + bI(g)$ for all $a, b \in \mathbb{R}$ and $f, g \in C[0,1]$, (2) $I(f) \geq 0$ whenever $f \geq 0$, and (3) $I(1) = 1$. Then, there exists a unique Borel probability measure μ on $[0,1]$ such that $I(f) = \int f d\mu$ for all $f \in C[0,1]$.

This shows that all positive linear functionals on $C[0,1]$ are given by Lebesgue integral with respect to a Borel measure. In other words, no need to go beyond the Lebesgue integral! The same result is true if we replace $[0,1]$ by any compact Hausdorff space. It is also true on a locally compact space (but then the linear functional is defined on the space of compactly supported continuous functions).

Remark 8

If you accept that positive linear functionals are natural things to consider, then Riesz's theorem associates to each of them a unique countably additive Borel probability measure. In other words, countable additivity is thrust on us, not imposed! In this sense, Riesz's representation theorem justifies the assumption of countable additivity in the definition of measure.

18. Lebesgue spaces

Fix a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. For $p > 0$, let V_p be the collection of all random variables $X : \Omega \to \mathbb{R}$ for which $\mathbf{E}[|X|^p] < \infty$. We also define V_∞ as the collection of all bounded random variables, i.e., all X for which there is a constant M such that $|X| \leq M$ $a.s.$

Claim 24

V_p is a vector space for any $0 < p \leq \infty$. Further, $V_p \supseteq V_q$ if $p \leq q$.

Proof. It is easy to see that V_∞ is a vector space. Indeed, if $|X| \leq M$ $a.s.$and $|Y| \leq M'$ $a.s.$, then $|\alpha X + \beta Y| \leq |\alpha|M + |\beta|M'$ $a.s.$

If $0 < p < \infty$, we recall that for any $x, y > 0$, we have $(x+y)^p \leq 2^{p-1}(x^p + y^p)$ if $p \geq 1$ and $(x+y)^p \leq x^p + y^p$ if $0 < p \leq 1$. Therefore, $|X+Y|^p \leq C_p(|X|^p + |Y|^p)$ where $C_p = 2^{p-1} \vee 1$. Thus, if $X, Y \in V_p$ then $X + Y \in V_p$. Further, if $X \in V_p$, then clearly $\alpha X \in V_p$ since $|\alpha X|^p \leq |\alpha|^p |X|^p$. This completes the proof that V_p is a vector space. This proves the first part of the claim.

Now suppose $p \leq q < \infty$. Then for any X, we have $|X|^p \leq |X|^q + 1$ (the extra 1 is needed for the case when $|X| < 1$). Using positivity of expectations, we get $\mathbf{E}[|X|^p] \leq 1 + \mathbf{E}[|X|^q]$. Hence, if $X \in V_q$ then $X \in V_p$. When $q = \infty$, this is even more obvious. ■

Next, we want to define a norm on V_p. To this end, we define $\|X\|_p := \mathbf{E}[|X|^p]^{\frac{1}{p}}$ for $X \in V_p$ for $p < \infty$ and $\|X\|_\infty := \inf\{t > 0 : |X| \leq t \ a.s.\}$. Then $\|tX\|_p = t\|X\|_p$ for $t > 0$, showing homogeneity. But there are issues with triangle inequality and strict positivity.

(1) Triangle inequality requires $\|X + Y\|_p \leq \|X\|_p + \|Y\|_p$ for any $X, Y \in V_p$. This is false if $p < 1$. Secondly, even for $p \geq 1$, this is not obvious to prove! We discuss it below.

(2) Strict positivity requires that $\|X\|_p = 0$ implies $X = 0$. But this is not true, as $\|X\|_p = 0$ if and only if $X = 0$ $a.s.$

Let us see how to deal with these issues.

Triangle inequality: As mentioned, triangle inequality fails for $p < 1$, even in the simplest non-trivial probability space!

Example 25

Let $\Omega = \{0, 1\}$ and $\mathbf{P}\{0\} = \mathbf{P}\{1\} = \frac{1}{2}$. Define $X(0) = a$, $X(1) = b$ and $Y(0) = b$, $Y(1) = a$ where $a, b > 0$. Then, $\|X\|_p = \|Y\|_p = (\frac{a^p+b^p}{2})^{\frac{1}{p}}$ while $\|X+Y\|_p = (a+b)$. Triangle inequality would imply that $(\frac{a+b}{2})^p \leq \frac{a^p+b^p}{2}$. But this is exactly the same as saying that $x \to x^p$ is a convex function, which is true if and only if $p \geq 1$.

Henceforth, we shall only take $p \geq 1$. But how does one prove Minkowski's inequality? We consider the important special cases of $p = 1, 2, \infty$ here. In the next section, we sketch a proof for general p.

(1) Case $p = 1$. In this case, since $|X + Y| \leq |X| + |Y|$, using positivity of expectation, we get

$$\|X + Y\|_1 = \mathbf{E}[|X + Y|] \leq \mathbf{E}[|X| + |Y|] = \mathbf{E}[|X|] + \mathbf{E}[|Y|] = \|X\|_1 + \|Y\|_1.$$

(2) Case $p = \infty$. If $|X| \leq M$ $a.s.$and $|Y| \leq M'$ $a.s.$, then $|X + Y| \leq M + M'$ $a.s.$Therefore, $\|X + Y\|_\infty \leq \|X\|_\infty + \|Y\|_\infty$.

(3) Case $p = 2$. The desired inequality is $\sqrt{\mathbf{E}[(X + Y)^2]} \leq \sqrt{\mathbf{E}[X^2]} + \sqrt{\mathbf{E}[Y^2]}$. Squaring and expanding $(X + Y)^2$, this reduces to $\mathbf{E}[XY] \leq \sqrt{\mathbf{E}[X^2]}\sqrt{\mathbf{E}[Y^2]}$. This inequality is indeed true, and is known as the Cauchy-Schwarz inequality.

The standard proof of Cauchy-Schwarz inequality is this: For any $t \in \mathbb{R}$, define $f(t) = \mathbf{E}[(X - tY)^2]$. By positivity of expectations, $f(t) \geq 0$, but also $f(t) = t^2\mathbf{E}[Y^2] - 2t\mathbf{E}[XY] + \mathbf{E}[X^2]$, a quadratic polynomial in t (assuming $\mathbf{E}[Y^2] \neq 0$). For this to be non-negative for

all t, we must have $(\mathbf{E}[XY])^2 \leq \mathbf{E}[X^2]\mathbf{E}[Y^2]$, proving Cauchy-Schwarz inequality and also showing that equality can hold if and only if X and Y are constant multiples of each other.

Strict positivity: Say that two random variables are equivalent and write $X \sim Y$ if $X = Y$ $a.s.$ If $X = X'$ $a.s.$and $Y = Y'$ $a.s.$, then $\alpha X + \beta X' = \alpha Y + \beta Y'$ $a.s.$Therefore, on the equivalence classes we can define addition and scalar multiplication (i.e., $\alpha[X] + [Y] = [\alpha X + Y]$ is a valid definition). In particular, if we restrict to V_p for some $p \geq 1$, then we get a vector space that we denote L^p (or $L^p(\Omega, \mathcal{F}, \mathbf{P})$ to describe the full setting). More precisely,

$$L^p = \{[X] : X \in V_p\}.$$

Then, L^p is a vector space, and $\|\cdot\|_p$ is a genuine norm on L^p (triangle inequality because $p \geq 1$ and strict positivity because we have quotiented by the equivalence relation).

Although elements of L^p spaces are equivalence classes of random variables, it is a standard abuse of language to speak of a random variable being in L^p, always keeping in mind that we don't distinguish two random variables that differ on a set of zero probability.

Completeness of L^p spaces: For $1 \leq p \leq \infty$, we have seen that $L^p(\Omega, \mathcal{F}, \mathbf{P})$ is a normed vector space. Automatically that makes it a metric space with distance defined by $\|X - Y\|_p$. The most important fact about L^p spaces (proof is left to measure theory class) is the following theorem of Riesz.

Theorem 25: Completeness of Lebesgue spaces [F. Riesz]

$L^p(\Omega, \mathcal{F}, \mathbf{P})$ is a complete metric space. That is, any Cauchy sequence converges.

This theorem is another indication that the Lebesgue integral is the right definition. For example, on the space $[0, 1]$, we could have define V_1 as the space of all Riemann integrable functions with norm defined by $\|f\| = \int_0^1 |f(t)|dt$. It would not be complete! An incomplete metric space may be thought of as missing many points which should have been there. In this sense, the L^p spaces define using Lebesgue integral has no missing points. Another indication that the Lebesgue integral is the right definition and needs no further improvement!

Remark 9: Banach and Hilbert spaces

A normed vector space that is complete as a metric space is called a *Banach space*. The space $L^p(\Omega, \mathcal{F}, \mathbf{P})$ and the space $C[0, 1]$ (with sup-norm) are prime examples of Banach spaces. The space L^2 alone is special in that its norm comes from an inner product. If $\langle X, Y \rangle = \mathbf{E}[XY]$, then by Cauchy-Schwars inequality, this is well defined for $X, Y \in L^2$ and defines

an inner product on L^2. Further, $\|X\|_2^2 = \langle X, X\rangle$. A Banach space whose norm comes from an inner product is called a *Hilbert space*. The space $L^2(\Omega, \mathcal{F}, \mathbf{P})$ is the rime (the only!) example of a Hilbert spaces. It is natural to ask if some of the other L^p spaces also have an inner product. The answer is no, since for any $p \neq 2$, the L^p-norm does not satisfy the parallelogram law: $\|X+Y\|^2 + \|X-Y\|^2 = 2\|X\|^2 + 2\|Y\|^2$ (see exercise below).

Exercise 17

On a two point probability space, construct random variables to show that parallelogram law fails for the L^p norm for $p \neq 2$.

19. Convex functions and Jensen's inequality

First we recall some basic facts about convex functions[4] on $\mathbb{R}$.

Definition 9

A function $\varphi : \mathbb{R} \mapsto \mathbb{R} \cup \{+\infty\}$ is said to be convex if $\varphi(x) < +\infty$ for some x and $\varphi(\alpha x + (1-\alpha)y) \leq \alpha\varphi(x) + (1-\alpha)\varphi(y)$ for all $x, y \in \mathbb{R}$ and any $0 \leq \alpha \leq 1$. Equivalently, we may say that the *epigraph* $E_\varphi := \{(x, y) \in \mathbb{R}^2 : y > \varphi(x)\}$ is a convex set (i.e., if two points are in the set, then the line segment joining them is contained in the set).

Example 26

Linear functions are convex. So is e^x. But $|x|^p$ is convex if and only if $p \geq 1$. If $\varphi(x) = 0 \times \mathbf{1}_{|x|\leq 1} + \infty \times \mathbf{1}_{|x|>1}$, then φ is convex. More generally, a convex function on an interval (defined exactly the same way as above) remains convex if extended as $+\infty$ outside the interval. Further, if φ_i, $i \in I$, are convex functions, so is $\sup_{i\in I} \varphi_i$.

Verifying that the above functions are indeed convex can be painful. A useful way to check that φ is convex is the following.

Exercise 18

If $\varphi : \mathbb{R} \mapsto \mathbb{R}$ is twice differentiable and $\varphi'' \geq 0$ everywhere, then φ is convex.

Let φ be convex and fix some $x_0 \in \mathbb{R}$ for which $\varphi(x_0) < \infty$. Define $D_\varphi(x, x_0) = \frac{\varphi(x)-\varphi(x_0)}{x-x_0}$ for $x \neq x_0$. This is the slope of the line segment joining $(x, \varphi(x))$ with $(x_0, \varphi(x_0))$ and could take the values $\pm\infty$.

[4] A good resource for a quick introduction to convex functions in one dimension is Rudin's *Real and Complex Analysis* (chapter 3)

Claim 26

$x \mapsto D_\varphi(x, x_0)$ is increasing on $\mathbb{R} \setminus \{x_0\}$.

Proof of the claim. Let $x < y$. To show $D_\varphi(x, x_0) \leq D_\varphi(y, x_0)$, we consider three cases, depending on which of x, y, x_0 is in the middle. We write out the proof for the case $x < y < x_0$, the other two being similar. In this case, after a simple rearrangement, $D_\varphi(x, x_0) \leq D_\varphi(y, x_0)$ is seen to be equivalent to

$$\varphi(y)(x_0 - x) \leq \varphi(x)(x_0 - y) + \varphi(x_0)(y - x).$$

This is true by the definition of convexity (since $y = \alpha x + (1-\alpha)x_0$ with $\alpha = \frac{x_0 - y}{x_0 - x}$). ■

The claim immediately implies the existence and finiteness of the left and right derivatives

$$\varphi'(x_0+) = \lim_{x \downarrow x_0} D_\varphi(x, x_0) \quad \text{and} \quad \varphi'(x_0-) = \lim_{x \uparrow x_0} D_\varphi(x, x_0)$$

and that $\varphi'(x_0-) \leq \varphi'(x_0+)$. In particular, φ is continuous at x_0. Further, if we choose $m \in \mathbb{R}$ so that $\varphi'(x_0-) \leq m \leq \varphi'(x_0+)$, then $D_\varphi(x, x_0) \leq m$ for $x < x_0$ and $D_\varphi(x, x_0) \geq m$ for $x \geq x_0$. Rearranging, this just says that $\varphi(x) \geq m(x - x_0) + \varphi(x_0)$ for all x_0. This last conclusion is called the *supporting hyperplane theorem* (it is valid in higher dimensions too). It can be stated as

"For any u with $\varphi(u) < \infty$, there is a line L_u in $\mathbb{R}^2$ that lies below the graph of φ and passes through $(u, \varphi(u))$. In particular, $\varphi(x) = \sup_u L_u(x)$."

As we saw in the examples above, a supremum of linear functions is convex. What we have proved here is the converse. See the discussion later for two other ways of arriving at this important conclusion. Now we state and prove Jensen's inequality.

Lemma 27: Jensen's inequality

Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a convex function. Let X be a r.v on some probability space. Assume that X and $\varphi(X)$ both have expectations. Then, $\varphi(\mathbf{E}X) \leq \mathbf{E}[\varphi(X)]$. The same assertion holds if φ is a convex function on some interval (a, b) and X takes values in (a, b) a.s.

Proof. Let $\mathbf{E}[X] = a$. Let $y = m(x - a) + \varphi(a)$ be the supporting line through $(a, \varphi(a))$. Since the line lies below the graph of φ, we have $m(X - a) + \varphi(a) \leq \varphi(X)$, $a.s.$ Take expectations to get $\varphi(a) \leq \mathbf{E}[\varphi(X)]$. ■

Supporting hyperplane theorem via Hahn-Banach theorem. If φ is a convex function and $\varphi(x_0) < \infty$, consider $A = \{(x, y) \in \mathbb{R}^2 : y > \varphi(x)\}$ and $B = \{(x_0, \varphi(x_0))\}$. Then A, B are disjoint convex sets and A is open. Hence there is a linear functional $L : \mathbb{R}^2 \mapsto \mathbb{R}$ and a number $d \in \mathbb{R}$ such that $L < d$ on A and $L \geq d$ on B.

On $\mathbb{R}^2$, linear functional are of the form $L(x,y) = ax + by$ but in our case we cannot have $b = 0$. Hence, writing $m = a/b$ and $c = d/b$, we see that $mx + c < y$ for $(x,y) \in A$ and $mx_0 + c \geq \varphi(x_0)$. Think for a moment to see that this is a supporting line at $(x_0, \varphi(x_0))$.

Supporting hyperplane via convex duality. For a convex function $\varphi : \mathbb{R} \mapsto \mathbb{R} \cup \{+\infty\}$, define its Legendre transform as $\varphi^* : \mathbb{R} \mapsto \mathbb{R} \cup \{+\infty\}$ as $\varphi^*(t) = \sup_x(tx - \varphi(x))$. Then φ^* is called the Legendre transform or the convex conjugate of φ. Here are the main facts.

Lemma 28

Let φ be a convex function. Then φ^* is a convex function and $(\varphi^*)^* = \varphi$.

We skip the proof for now (if I get time later I shall write this). But the point is that $\varphi^{**} = \varphi$ means that $\varphi(x) = \sup_t xt - \varphi^*(t)$. For each t, the function $x \mapsto xt - \varphi^*(t)$ is linear, hence this gives a representation of φ as a supremum of linear functions. That is the import of the supporting hyperplane theorem.

Proof of Lemma 28. By definition of the conjugate, $\varphi^*(t) \geq tx - \varphi(x)$ for any x, t. Rewrite this as $\varphi(x) \geq tx - \varphi^*(t)$ and take supremum over t to get $\varphi(x) \geq \varphi^{**}(x)$. ■

Legendre transformation is quite fundamental and appears in many contexts (for example, the Lagrangian and Hamiltonian in classical mechanics are convex conjugates of each other). Here is an important example from mathematics.

Example 27

Let $1 < p < \infty$ and let $\varphi(x) = \frac{1}{p}|x|^p$. To compute φ^*, first take $t > 0$ and observe that $tx - \varphi(x)$ is negative for $x < 0$, hence the supremum is attained for some positive x. Writing $tx - \frac{1}{p}x^p$ for $x > 0$, elementary calculus shows that the maximizer satisfies $x^{p-1} = t$ and the maximum value is $(1 - \frac{1}{p})t^{p/(p-1)}$. A similar calculation works for $t < 0$. Thus if $q = p/(p-1)$, then $\varphi^*(t) = \frac{1}{q}|t|^q$. Observe that this q is the number that satisfies $\frac{1}{p} + \frac{1}{q}$, a relationship familiar to us in functional analysis. This relationship between conjugate exponents is the reason why L^q is the dual of L^p, etc.

Exercise 19

What happens when $p = 1$?

Exercise 20

For a convex function $\varphi : \mathbb{R}^d \mapsto \mathbb{R} \cup \{\infty\}$, the Legendre transform is defined as $\varphi^*(t) = \sup_{x \in \mathbb{R}^d} \langle x, t\rangle - \varphi(x)$. The lemma above remains valid.
If $x \mapsto \varphi(x)$ is a norm on $\mathbb{R}^d$, show that the dual norm is given by $t \mapsto \varphi^*(t)$. In particular, the ℓ^p norm has dual norm ℓ^q.

20. Further inequalities for expectation

We gave a proof of Minkowski's inequality for $p = 1, 2, \infty$ in the previous section. Now we prove it for all $p \geq 1$.

Lemma 29: Minkowski's inequality

For any $p \geq 1$, we have $\|X + Y\|_p \leq \|X\|_p + \|Y\|_p$.

Proof. The case $p = \infty$ was proved earlier, so take $1 \leq p < \infty$. Let $X' = X/\|X\|_p$ and $Y' = Y/\|Y\|_p$. By convexity of $x \to x^p$, we see that $|aX' + bY'|^p \leq a|X'|^p + b|Y'|^p$ where $a = \frac{\|X\|_p}{\|X\|_p + \|Y\|_p}$ and $b = \frac{\|Y\|_p}{\|X\|_p + \|Y\|_p}$. Take expectations and observe that $\mathbf{E}[|aX' + bY'|^p] = \frac{\mathbf{E}[|X+Y|^p]}{(\|X\|_p + \|Y\|_p)^p}$ while $\mathbf{E}[a|X'|^p + b|Y'|^p] = 1$ since $\mathbf{E}[|X'|^p] = \mathbf{E}[|Y'|^p] = 1$. Thus we get

$$\frac{\mathbf{E}[|X + Y|^p]}{(\|X\|_p + \|Y\|_p)^p} \leq 1,$$

which is precisely Minkowski's inequality. ■

Lastly, we prove Hölder's inequality of which the most important special case is the Cauchy-Schwarz inequality.

Lemma 30: Cauchy-Schwarz and Hölder inequalities

(1) If X, Y are L^2 random variables on a probability space, then XY is integrable and $\mathbf{E}[XY]^2 \leq \mathbf{E}[X^2]\mathbf{E}[Y^2]$.

(2) If X, Y are L^p r.v.s on a probability space, then for any $p, q \geq 1$ satisfying $p^{-1} + q^{-1} = 1$, we have $XY \in L^1$ and $\|XY\|_1 \leq \|X\|_p \|Y\|_q$.

Proof. Cauchy-Schwarz is a special case of Hölder with $p = q = 2$ (we also gave a direct proof in the previous section). Hölder's inequality follows by applying the inequality $a^p/p + b^q/q \geq ab$ valid for $a, b \geq 0$, to $a = |X|/\|X\|_p$ and $b = |Y|/\|Y\|_q$ and taking expectations.

The inequality $a^p/p + b^q/q \geq ab$ is evident by noticing that the rectangle $[0, a] \times [0, b]$ (with area ab) is contained in the union of the region$\{(x, y) : 0 \leq x \leq a,\ 0 \leq y \leq x^{p-1}\}$ (with area a^p/p) and the region $\{(x, y) : 0 \leq y \leq b,\ 0 \leq x \leq y^{q-1}\}$ (with area b^q/q). This is because the latter regions are

the regions between the x and y axes (resp.) and curve $y = x^{p-1}$ which is also the curve $x = y^{q-1}$ since $(p-1)(q-1) = 1$. ■

Remark 10

To see the role of convexity, here is another way to prove that $a^p/p + b^q/q \geq ab$. Set $a' = p \log a$ and $b' = q \log b$ and observe that the desired inequality is equivalent to $\frac{1}{p}e^{a'} + \frac{1}{q}e^{b'} \geq e^{\frac{1}{p}a' + \frac{1}{q}b'}$, which follows from the convexity of $x \to e^x$.

In the study of L^p spaces, there is a close relationship between L^p and L^q where $\frac{1}{p} + \frac{1}{q} = 1$. In the proof of Hölder's inequality, we see one elementary way in which it arises (the inverse of $y = x^{p-1}$ is $x = y^{q-1}$). Another big-picture description is via the convex duality that we mentioned earlier.

21. Change of variables

Lemma 31

Let $T : (\Omega_1, \mathcal{F}_1, \mathbf{P}) \to (\Omega_2, \mathcal{F}_2, \mathbf{Q})$ be measurable and $\mathbf{Q} = \mathbf{P}T^{-1}$. If X is an integrable r.v. on Ω_2, then $X \circ T$ is an integrable r.v. on Ω_1 and $\mathbf{E}_{\mathbf{P}}[X \circ T] = \mathbf{E}_{\mathbf{Q}}[X]$.

Proof. For a simple r.v., $X = \sum_{i=1}^n c_i \mathbf{1}_{A_i}$, where $A_i \in \mathcal{F}_2$, it is easy to see that $X \circ T = \sum_{i=1}^n c_i \mathbf{1}_{T^{-1}A_i}$ and by definition $\mathbf{E}_{\mathbf{P}}[X \circ T] = \sum_{i=1}^n c_i \mathbf{P}\{T^{-1}A_i\} = \sum_{i=1}^n c_i \mathbf{Q}\{A_i\}$ which is precisely $\mathbf{E}_{\mathbf{Q}}[X]$. Use MCT to get to positive r.v.s and then to general integrable r.v.s. ■

Corollary 32

Let X_i, $i \leq n$, be random variables on a common probability space. Then for any Borel measurable $f : \mathbb{R}^n \to \mathbb{R}$, the value of $\mathbf{E}[f(X_1, \ldots, X_n)]$ (if it exists) depends only on the joint distribution of $X_1, \ldots X_n$.

Proof. Consider $T = (X_1, \ldots, X_n) : \Omega \to \mathbb{R}^n$. Then $\mu := \mathbf{P} \circ T^{-1}$ is (by definition) the joint distribution of $X_1, \ldots, X_n$. The Lemma gives $\mathbf{E}_{\mathbf{P}}[f(X_1, \ldots, X_n)] = \int_{\mathbb{R}^n} f(t) d\mu(t)$. Clearly, the right hand side depends only on the measure μ. ■

Remark 11

The change of variable result shows the irrelevance of the underlying probability space to much of what we do. In any particular situation, all our questions may be about a finite or infinite collection of random variables X_i. Then, the answers depend only on the joint distribution of these random variables and not any other details of the underlying probability space. For instance, we can unambiguously talk of the expected value of an Exp(λ) random variable, the value being $1/\lambda$ regardless of the details of the probability space on

which the random variable is defined. Thus, statements in theorems and problems go like "Let $X_1, \ldots, X_n$ be random variables with a multivariate normal distribution with mean and variance..." without bothering to say what the probability space is.

Change of variable formula for densities: We discuss densities more in the next section, but for now consider a Borel probability measure μ on $\mathbb{R}^n$. We say that it has a density function $f : \mathbb{R}^n \to \mathbb{R}_+$ if f is a Borel measurable function and $\mu(A) = \int_A f(x)dm(x)$ where m is the Lebesgue measure on $\mathbb{R}^n$. Here, $\int_A f(x)dm(x)$ is just the notation for $\int_{\mathbb{R}^n} f(x)\mathbf{1}_A(x)dm(x)$. Strictly speaking, we have define Lebesgue integral only for probability measures (m is not a probability measure), but a similar procedure constructs Lebesgue integral with respect to general measures.

Now consider a transformation $T : \mathbb{R}^n \to \mathbb{R}^n$ and let $\nu = \mu \circ T^{-1}$ where μ is a probability measure with density f. In case T is nice enough, the change of variable formula shows that ν also has a density and gives a recipe for finding it in terms of f and T. We omit the proof.

Proposition 33

Let U, V be open subsets of $\mathbb{R}^n$ and let $T : U \to V$ be a bijective smooth function such that $T^{-1} : V \to U$ is also smooth. Let X be a random vector on some probability space, taking values in U and assume that its distribution has density f with respect to Lebesgue measure on U. Let $Y = T(X)$, so that Y takes values in V. Then, Y has density g with respect to Lebesgue measure on V where $g(x) = f(T^{-1}x)|\det J[T^{-1}](x)|$.

21.1. **Distribution of the sum, product etc.** Whenever $Y = T(X)$, in principle we can find the distribution of Y from the distribution of X (just push forward under T). However, in practise it may be very hard to actually compute. The usefulness of the change of variable formula for densities is that, in some situations, the density of Y can be found from the density of X. In particular, it is important to know how to compute the distribution of the sum or product of two random variables, given their joint distribution.

Example 28

Suppose (X_1, X_2) has density $f(x, y) = e^{-x-y}$ on $\mathbb{R}^2_+$. How to find the distribution of $X_1 + X_2$?

Define $T(x_1, x_2) = (x_1 + x_2, x_2)$. Then T is a bijection from $\mathbb{R}^2_+$ onto $V = \{(u, v) : u > v > 0\}$ and $T^{-1}(u, v) = (u - v, v)$. The Jacobian determinant is found to be 1. Hence, the density of $(Y_1, Y_2) = T(X_1, X_2)$ is given by $g(u, v) = f(u - v, v)\mathbf{1}_{u>v>0} = e^{-u}\mathbf{1}_{u>v>0}$. This gives

the joint density of (Y_1, Y_2). We can get the density of Y_1 by integrating out v. We get $\int_0^u e^{-u} dv = ue^{-u}$. This is the Gamma$(2, 1)$ density.

Actually there are a couple of facts that we have invoked without comment in this example and in examples to come below. We computed the joint density of (Y_1, Y_2) to be $g(u, v)$. What this means is that $\mathbf{P}\{(Y_1, Y_2) \in R_{(a,b)}\} = \int_{R_{a,b}} g(y)dm(y)$ where $R_{(a,b)} = (-\infty, a] \times (-\infty, b]$. From this, we conclude that the density of Y_1 is $h(a) = \int_{-\infty}^{\infty} g(a, v)dv$. In doing this, we are implicitly using the fact that a multiple integral is the same as an iterated integral. You have probably seen this in Analysis class for Riemann integral. A much better result for Lebesgue integrals will come in a later section under the name of *Fubini's theorem*.

A few useful transformations are covered below.

Example 29

Suppose (X, Y) has density $f(x, y)$ on $\mathbb{R}^2$.

(1) X has density $f_1(x) = \int_{\mathbb{R}} f(x, y)dy$ and Y has density $f_2(y) = \int_{\mathbb{R}} f(x, y)dx$. This is because, for any $a < b$, we have

$$\mathbf{P}(X \in [a, b]) = \mathbf{P}((X, Y) \in [a, b] \times \mathbb{R}) = \int\int_{[a,b]\times\mathbb{R}} f(x, y)dxdy = \int_{[a,b]} \left(\int_{\mathbb{R}} f(x, y)dy \right) dx.$$

This shows that the density of X is indeed f_1.

(2) Density of X^2 is $\left(f_1(\sqrt{x}) + f_1(-\sqrt{x})\right)/2\sqrt{x}$ for $x > 0$. Here we notice that T is one-one on $\{x > 0\}$ and $\{x < 0\}$ (and $\{x = 0\}$ has zero measure under f), so the change of variable formula is used separately for the two domains and the result is added.

(3) The density of $X + Y$ is $g(t) = \int_{\mathbb{R}} f(t - v, v)dv$. To see this, let $U = X + Y$ and $V = Y$. Then the transformation is $T(x, y) = (x+y, y)$. Clearly $T^{-1}(u, v) = (u-v, v)$ whose Jacobian determinant is 1. Hence by Proposition **??**, we see that (U, V) has the density $g(u, v) = f(u - v, v)$. Now the density of U can be obtained like before as $h(u) = \int g(u, v)dv = \int f(u - v, v)dv$.

(4) To get the density of XY, we define $(U, V) = (XY, Y)$ so that for $v \neq 0$, we have $T^{-1}(u, v) = (u/v, v)$ which has Jacobian determinant v^{-1}.

Exercise 21

(1) Suppose (X, Y) has a continuous density $f(x, y)$. Find the density of X/Y. Apply to the case when (X, Y) has the *standard bivariate normal distribution* with density $f(x, y) = (2\pi)^{-1} \exp\{-\frac{x^2+y^2}{2}\}$ and show that X/Y has Cauchy distribution.

(2) Find the distribution of $X + Y$ if (X, Y) has the standard bivariate normal distribution.

(3) Let $U = \min\{X, Y\}$ and $V = \max\{X, Y\}$. Find the density of (U, V).

22. Absolute continuity and singularity

Consider a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. Let $X : \Omega \to \mathbb{R}$ be a non-negative random variable with $\mathbf{E}[X] = 1$. Define $\mathbf{Q}(A) = \mathbf{E}[X\mathbf{1}_A]$ for $A \in \mathcal{F}$. Then, $\mathbf{Q}$ is a probability measure on $(\Omega, \mathcal{F})$. The only non-trivial thing to check is that if $A_n, A \in \mathcal{F}$ and $A_n \uparrow A$ then $\mathbf{Q}(A_n) \uparrow \mathbf{Q}(A)$. This follows from MCT, since $X\mathbf{1}_{A_n} \uparrow X\mathbf{1}_A$. All this clearly remains valid even if $\mathbf{P}$ and $\mathbb{Q}$ were infinite measures and X is a general non-negative measurable function.

If two measures μ, ν (not necessarily probability measures) on $(\Omega, \mathcal{F})$ are such that $\nu(A) = \int_A f d\mu$ for all $A \in \mathcal{F}$ (where $\int_A f d\mu$ is just the notation for the Lebesgue integral of $f\mathbf{1}_A$ with respect to μ) for some non-negative measurable function f, then we say that ν has density f with respect to μ.

Question: Given two measures μ, ν on $(\Omega, \mathcal{F})$, does ν have a density with respect to μ and is it unique?

The uniqueness part is easy. If a density exists, it is unique (in $L^1(\mu)$). Indeed, if $\nu(A) = \int_A f d\mu = \int_A g d\mu$ for some f, g, then $h := f - g$ satisfies $\int_A h d\mu = 0$ for all $A \in \mathcal{F}$. Take $A = \{h > 0\}$ to get $\int h\mathbf{1}_{h>0} d\mu = 0$. But $h\mathbf{1}_{h>0}$ is a non-negative measurable function, hence it must be that $h\mathbf{1}_{h>0} = 0$ $a.s.[\mu]$. This implies that $\mu\{h > 0\} = 0$. Similarly $\mu\{h < 0\} = 0$ and we see that $h = 0$ $a.s.[\mu]$ or equivalently $f = g$ $a.s[\mu]$. The density is unique up to sets of μ-measure zero. More than that cannot be asked because, if f is a density and $g = f$ $a.s.[\mu]$, then it follows that $\int_A g d\mu = \int_A f d\mu$ and hence g is also a density of ν with respect to μ.

Existence of density is a more subtle question. First let us see some examples.

Example 30

On $([0, 1], \mathcal{B}, \lambda)$ let ν be the measure with distribution $F_\nu(x) = x^2$. Then ν has density $f(x) = 2x\mathbf{1}_{x\in[0,1]}$ with respect to λ. Indeed, if we set $\theta(A) = \int_A f d\lambda$, then θ and ν are two

measures on $[0,1]$ that agree on all intervals, since $\int_{[a,b]} f d\lambda = b^2 - a^2$ for any $[a,b] \subseteq [0,1]$. By the $\pi - \lambda$ theorem, $\theta = \nu$.

Note that the same logic works whenever $\nu \in \mathcal{P}(\mathbb{R})$ and F_ν has a continuous (or piecewise continuous) derivative. If $f = F_\nu'$, by the fundamental theorem of Calculus, $\int_{[a,b]} f d\lambda = F_\nu(b) - F_\nu(a)$ and hence by the same reasoning as above, ν has density f with respect to Lebesgue measure.

Example 31

Let Ω be some set and let $a_1, \ldots, a_n$ be distinct elements in Ω. Let $\nu = \sum_{k=1}^n p_k \delta_{a_k}$ and let $\mu = \sum_{k=1} q_k \delta_{a_k}$ where p_i, q_i are non-negative numbers such that $\sum_i p_i = \sum_i q_i = 1$.

Assume that $q_i > 0$ for all $i \leq n$. Then define $f(x) = \frac{p_i}{q_i}$ for $x = a_i$ and in an arbitrary fashion for all other $x \in \Omega$. Then, f is the density of ν with respect to μ. The key point is that $\int f \mathbf{1}_{\{a_i\}} d\mu = f(a_i)\mu\{a_i\} = p_i = \nu\{a_i\}$.

On the other hand, if $q_i = 0 < p_i$ for some i, then ν cannot have a density with respect to μ (why?).

Let us return to the general question of existence of density of a measure ν with respect to a measure μ (both measures are defined on $(\Omega, \mathcal{F})$). As in the last example, there is one necessary condition for the existence of density. If $\nu(A) = \int f \mathbf{1}_A d\mu$ for all A, then if $\mu(A) = 0$ we must have $\nu(A) = 0$ (since $f\mathbf{1}_A = 0$ $a.s[\mu]$). In other words, if there is even one set $A \in \mathcal{F}$ such that $\nu(A) > 0 = \mu(A)$, then ν cannot have a density with respect to μ. Let us make a definition.

Definition 10

Two measures μ and ν on the same $(\Omega, \mathcal{F})$ are said to be *mutually singular* and write $\mu \perp \nu$ if there is a set $A \in \mathcal{F}$ such that $\mu(A) = 0$ and $\nu(A^c) = 0$. We say that μ is *absolutely continuous to* ν and write $\mu \ll \nu$ if $\mu(A) = 0$ whenever $\nu(A) = 0$.

Remark 12

(1) Singularity is a symmetric relation, absolute continuity is not. If $\mu \ll \nu$ and $\nu \ll \mu$, then we say that μ and ν are *mutually absolutely continuous*. (2) If $\mu \perp \nu$, then we cannot also have $\mu \ll \nu$ (unless $\mu = 0$). (3) Given μ and ν, it is not necessary that they be singular or absolutely continuous to one another. (4) Singularity is not reflexive but absolute continuity is. That is, $\mu \ll \mu$ but μ is never singular to itself (unless μ is the zero measure).

Example 32

Uniform($[0,1]$) and Uniform($[1,2]$) are singular. Uniform($[1,3]$) is neither absolutely continuous nor singular to Uniform($[2,4]$). Uniform($[1,2]$) is absolutely continuous with respect to Uniform($[0,4]$) but not conversely. All these uniforms are absolutely continuous to Lebesgue measure. Any measure on the line that has an atom (eg., δ_0) is not absolutely continuous to Lebesgue measure. A measure that is purely discrete is singular with respect to Lebesgue measure. A probability measure on the line with density (eg., $N(0,1)$) is absolutely continuous to λ. In fact $N(0,1)$ and λ are mutually absolutely continuous. However, the exponential distribution is absolutely continuous to Lebesgue measure, but not conversely (since $(-\infty, 0)$, has zero probability under the exponential distribution but has positive Lebesgue measure).

Returning to the existence of density, we saw that for ν to have a density with respect to μ, it is necessary that $\nu \ll \mu$. This condition is also sufficient!

Theorem 34: Radon Nikodym theorem

Suppose μ and ν are two finite measures on $(\Omega, \mathcal{F})$. If $\nu \ll \mu$, then ν has a density with respect to μ.

A first attempt at proof: Let $H = L^2(\mu)$ and define $L : H \mapsto \mathbb{R}$ by $Lf = \int f d\nu$. Suppose we could show that L is well-defined (then it is clearly linear) and bounded, i.e., $|Lf| \leq C\|f\|_H$ for all $f \in H$. Then, by the Riesz representation theorem for linear functionals on a Hilbert space, it follows that $Lf = \langle f, \psi \rangle$ for some $\psi \in H$. Take $f = \mathbf{1}_A$ with $A \in \mathcal{F}$ to see that $\nu(A) = \int_A \psi \, d\mu$. This is what we want to show.

The problem is that L need not be bounded. Indeed, it it were true, the above argument would have shown that the Radon -Nikodym derivative of ν w.r.t. μ is in $L^2(\mu)$, which is false in general! For example, let $\nu(A) = \int_A \frac{1}{\sqrt{x}} d\lambda(x)$, where λ is the Lebesgue measure on $[0,1]$. Then the Radon-Nikodym derivative is $1/\sqrt{x}$, whose square is not integrable w.r.t. μ. The proof below overcomes this issue by a small trick.

Proof of the Radon Nikodym theorem. Let $\theta = \mu + \nu$ and let $H = L^2(\Omega, \mathcal{F}, \theta)$. Define $L : H \mapsto \mathbb{R}$ by $Lf = \int f d\nu$. Since (note that $\int g d\nu \leq \int g d\theta$ for any $g \geq 0$)

$$\left| \int f d\nu \right| \leq \int |f| d\nu \leq \int |f| d\theta \leq \sqrt{\theta(\Omega)} \left(\int |f|^2 d\theta \right)^{\frac{1}{2}},$$

it follows that L is well-defined and $|Lf| \leq C\|f\|_H$ with $C = \sqrt{\theta(\Omega)}$.Therefore, L is bounded and $Lf = \int f\varphi \, d\theta$ for some $\varphi \in H$. Rewrite this as

$$\int f(1-\varphi)d\nu = \int f\varphi d\mu \quad \text{for all } f \in H. \tag{1}$$

From this identity, it is clear that $0 \leq \varphi \leq 1$ $a.s.[\mu]$ (hence also $a.s.[\nu]$). Further, setting $f = \mathbf{1}_{\varphi=1}$, we see that the left hand side is zero while the right hand side is $\mu\{\varphi = 1\}$. Thus, $\varphi < 1$ $a.s.[\mu]$ (hence also $a.s.[\nu]$).

Now for any $A \in \mathcal{F}$ and $\delta > 0$, setting $f = \frac{1}{1-\varphi}\mathbf{1}_A\mathbf{1}_{\varphi\leq 1-\delta}$ (which is bounded above by $1/(1-\delta)$ and hence in H), we get that $\nu(A\cap\{\varphi \leq 1-\delta\}) = \int_A \psi\mathbf{1}_{\varphi\leq 1-\delta}d\mu$, where $\psi = \varphi/(1-\varphi)$. Set $\delta = 1/n$ and let $n \uparrow \infty$. We get $\nu(A \cap \{\varphi < 1\}) = \int \psi\mathbf{1}_{\varphi<1}d\mu$. Since $\varphi < 1$ almost surely with respect to both measures, it is redundant to write that, and we get $\nu(A) = \int_A \psi d\mu$. ■

Exercise 22: Lebesgue decomposition

Let μ, ν be two finite measures on $(\Omega, \mathcal{F})$. Show that we can write $\nu = \nu_1 + \nu_2$, where ν_1, ν_2 are measures on $\mathcal{F}$ and $\nu_1 \ll \mu$ and $\nu_2 \perp \mu$. This decomposition is unique. [*Hint:* Follow the steps in the proof of Radon-Nikodym theorem and consider theset $\{\varphi = 1\}$ carefully!]

23. Some singular probability measures

This section is not directly needed for what comes next in the course. But these are some natural directions suggested by the previous discussion of absolute continuity and singularity of measures.

Is there any $\mu \in \mathcal{P}(\mathbb{R})$ that is singular to Lebesgue measure on $\mathbb{R}$? Of course, any discrete probability measure is singular, since it gives probability one to a countable set while Lebesgue measure gives probability zero to that set. The interesting question is whether there is a singular μ that has no atoms. For this, we must spread our set on some uncountable set of zero Lebesgue measure. The first example that comes to mind is the standard Cantor set.

Recall that the middle-thirds Cantor set is defined as the decreasing intersection K of K_ns where $K_0 = [0,1]$, $K_1 = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$, $K_3 = [0, \frac{1}{9}] \cup [\frac{2}{9}, \frac{3}{9}] \cup [\frac{6}{9}, \frac{7}{9}] \cup [\frac{8}{9}, 1]$, and so on. In general, K_n is a union of 2^n intervals each of length 3^{-n}, and K_{n+1} is got from K_n by deleting the middle third open subinterval of each of these intervals. An alternate description of the Cantor set is

$$K = \left\{ x \in [0,1] : x = \sum_{n=1}^{\infty} \frac{x_n}{3^n} \text{ for some } x_n \in \{0, 2\} \right\}.$$

In other words, it consists of those numbers that have a ternary (base-3) expansion without using the digit 1.

Example 33: Cantor measure

Let K be the middle-thirds Cantor set. Consider the canonical probability space $([0,1], \mathcal{B}, \lambda)$ and the random variable $X(\omega) = \sum_{k=1}^{\infty} \frac{2B_k(\omega)}{3^k}$, where $B_k(\omega)$ is the kth binary digit of ω (i.e., $\omega = \sum_{k=1}^{\infty} \frac{B_k(\omega)}{2^k}$). Then X is measurable (we saw this before). Let $\mu := \lambda \circ X^{-1}$ be the pushforward measure.

Then, $\mu(K) = 1$, because X takes values in numbers whose ternary expansion has no ones. Further, for any $t \in K$, $X^{-1}\{t\}$ is a set with atmost two points and hence $\mu\{t\} = 0$. Thus μ has no atoms and must have a continuous CDF. Since $\mu(K) = 1$ but $\lambda(K) = 0$, we also see that $\mu \perp \lambda$.

Exercise 23: Alternate construction of Cantor measure

Write $K = \cap K_n$ as in the definition of the Cantor set. Let μ_n be the uniform probability measure on K_n, i.e., $\mu_n(A) = (3/2)^n \lambda(A \cap K_n)$ for all $A \in \mathcal{B}_{\mathbb{R}}$. Show that F_{μ_n}s converge uniformly to a CDF F and that the measure having this CDF is the Cantor measure constructed above.

Example 34: Bernoulli convolutions - a fun digression (omit if unclear!)

We generalize the previous example. For any $\alpha > 1$, define $X_\alpha : [0,1] \to \mathbb{R}$ by $X_\alpha(\omega) = \sum_{k=1}^{\infty} \alpha^{-k} B_k(\omega)$. Let $\mu_\alpha = \lambda \circ X_\lambda^{-1}$ (did you check that X_α is measurable?). These measures are called Bernoulli convolutions. For $\alpha = 3$, this is almost the same as $1/3$-Cantor measure, except that we have left out the irrelevant factor of 2 (so μ_3 is a probability measure on $\frac{1}{2}K := \{x/2 : x \in K\}$) and hence is singular. For $\alpha = 2$, the map X_α is identity, and hence μ_2 is the Lebesgue measure on $[0,1]$, certainly absolutely continuous to Lebsegue measure. What about the singularity and absolute continuity of μ_α for other values of α?

Exercise 24

For any $\alpha > 2$, show that μ_α is singular w.r.t. Lebesgue measure.

Hence, one might expect that μ_α is absolutely continuous to Lebesgue measure for $1 < \alpha < 2$. This is false! Paul Erdős showed that μ_α is singular to Lebesgue measure whenever α is a Pisot-Vijayaraghavan number, i.e., if α is an algebraic number all of whose conjugates have modulus less than one!! It is an open question as to whether these are the only exceptions.

23.1. **Hausdorff measures.** Consider two Cantor type sets: A consisting of those numbers who decimal expansion does not have the disgit 5 and B consisting of those numbers who decimal expansion does not have any odd digit. Both have Lebesgue measure zero. Is there another

measure that can measure the sizes of these sets (one might feel that B is somehow smaller than A, but in what sense?).

Let (X, d) be a compact metric space. Fix $\alpha > 0$ and define for any $A \subseteq X$,

$$H^*_\alpha(A) = \inf \left\{ \sum_{n=1}^{\infty} \text{dia}(B_n)^\alpha : B_n \text{ are open balls whose union covers } A \right\}.$$

It is easy to check that $H^*_\alpha(A) \leq H^*_\alpha(B)$ if $A \subseteq B$ and $H^*_\alpha(\cup_n A_n) \leq \sum_n H^*_\alpha(A_n)$. Thus H^*_α is an outer measure H_α and can be used to construct a measure on $(X, \mathcal{B}_X)$ (one must check many things, for example that the Caratheodary construction gives a sigma algebra containing all Borel sets). As it happens, for most α, the measure H_α turns out to be trivial. For example, if $X = [0, 1]$, then for any interval I, one can check that $H_\alpha(I) = 0$ if $\alpha > 1$ and $H_\alpha(I) = \infty$ if $\alpha < 1$. For $\alpha = 1$, we get the Lebesgue measure.

For a general X, again there is always a value α_0 such that for any open ball B we have $H_\alpha(B) = 0$ if $\alpha > \alpha_0$ and $H_\alpha(B) = \infty$ if $\alpha < \alpha_0$. At $\alpha = \alpha_0$, we may or may not get a meaningful measure. If we do, then H_{α_0} is called the *Hausdorff measure* on X. Whether H_{α_0} is trivial or not, the number α_0 is called the *Hausdorff dimension* of X.

Example 35

Let $X = K$, the middle-thirds Cantor set. Then $\alpha_0 = \log 2 / \log 3$ and H_{α_0} is precisely the Cantor measure that we constructed earlier.

24. Conditional probability and expectation - a first view

So far (and for a few lectures next), we have seen how a rigorous framework for probability theory is provided by measure theory. We have not yet touched the two most important concepts in probability, *independence* and *conditional probability*. We shall see independence very shortly but may not have time to study conditional probability in detail in this course. But one of the important aspects of Kolmogorov's axiomatization of probability using measure theory was to define conditional probability using the Radon-Nikodym theorem. Here is a teaser for that story.

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. Let X be a random variable that takes finitely many values $a_1, \ldots, a_n$ with $\mathbf{P}\{X = a_k\} > 0$ for each k. Then, the law of total probability says that for any $A \in \mathcal{F}$,

$$\mathbf{P}(A) = \sum_{k=1}^{n} \mathbf{P}(A \,\big|\, X = a_k)\mathbf{P}\{X = a_k\}$$

where $\mathbf{P}(A \,\big|\, X = a_k) = \frac{\mathbf{P}\{A \cap \{X = a_k\}\}}{\mathbf{P}\{X = k\}}$. Now suppose X takes uncountably many values, for eg., X has density f_X. Then, we would like to write

$$\mathbf{P}(A) = \int \mathbf{P}(A \,\big|\, X = t) f_X(t) dt$$

where f_X is the density of X and perhaps even generalize it to the case when X does not have density as $\mathbf{P}(A) = \int \mathbf{P}(A \mid X = t) d\mu_X(t)$. The question is, what is $\mathbf{P}(A \mid X = t)$? The usual definition makes no sense since $\mathbf{P}\{X = t\} = 0$.

The way around is this. Fix $A \in \mathcal{F}$ and set $\nu_A(I) = \mathbf{P}\{A \cap \{X \in I\}\}$ for $I \in \mathcal{B}_{\mathbb{R}}$. Then ν is a Borel probability measure on as a measure on $\mathbb{R}$. If μ_X is the distribution of X, then clearly $\nu_A \ll \mu_X$ (if $\mu_X(I) = 0$ then $\mathbf{P}\{X \in I\} = 0$ which clearly implies that $\nu_A(I) = 0$). Hence, by the Radon-Nikodym theorem, ν_A has a density $f_A(t)$ with respect to μ_X. In other words,

$$\mathbf{P}(A \cap \{X \in I\}) = \int_I f_A(t) d\mu_X(t)$$

and in particular, $\mathbf{P}(A) = \int_{\mathbb{R}} f_A(t) d\mu_X(t)$. Then, we may *define* $f_A(t)$ as the conditional probability of A given $X = t$! Note that f_A is defined only almost everywhere, hence $\mathbf{P}(A \mid X = t)$ should also be interpreted as being defined for almost every t (w.r.t. μ_X). This way, the intuitive notion of conditional probability is brought into the ambit of measure theoretical probability. We now elaborate on this a bit.

Let $\mathbf{P}, \mathbf{Q}$ be probability measures on $(\Omega, \mathcal{F})$. Assume that $\mathbf{Q} \ll \mathbf{P}$. Then there is a $X \in L^1(\Omega, \mathcal{F}, \mathbf{P})$ such that

$$\mathbf{Q}(A) = \int_A X d\mathbf{P} \quad \text{for all } A \in \mathcal{F}.$$

Now suppose $\mathcal{G} \subseteq \mathcal{F}$ is a sub-sigma algebra. Let $\mathbf{P}', \mathbf{Q}'$ be the restrictions of $\mathbf{P}, \mathbf{Q}$ to $\mathcal{G}$. It is trivially the case that $\mathbf{Q}' \ll \mathbf{P}'$. Hence, again by the Radon-Nikodym theorem, there is some $X' \in L^1(\Omega, \mathcal{G}, \mathbf{P}')$ such that $\mathbf{Q}'(A) = \int_A X' d\mathbf{P}'$ for all $A \in \mathcal{G}$. The last statement can also be written as

$$\mathbf{Q}(A) = \int_A X' d\mathbf{P} \quad \text{for all } A \in \mathcal{G}.$$

This X' is not the same as X, because the latter need not be $\mathcal{G}$-measurable.

Now start with any integrable random variable Y on $(\Omega, \mathcal{F}, \mathbf{P})$. Writing as $Y_+ - Y_-$ and applying the above steps to find Y'_+, Y'_- (these are $\mathcal{G}$-measurable and give the same integrals as Y_+, Y_- over sets in $\mathcal{G}$). Writing $Y' = Y'_+ - Y'_-$, we have shown that there is a $\mathcal{G}$-measurable random variable Y' such that

$$\int_A Y \, d\mathbf{P} = \int_A Y' \, d\mathbf{P} \quad \text{for all } A \in \mathcal{G}.$$

This Y' is called the conditional expectation of Y w.r.t. $\mathcal{G}$ and denoted $\mathbf{E}[Y \mid \mathcal{G}]$.

Example 36

Again consider $(\Omega, \mathcal{F}, \mathbf{P})$ and a measurable partition $\{A_1, \ldots, A_k\}$ with $\mathbf{P}(A_i) > 0$ for all i. Let $\mathcal{G} = \sigma\{A_1, \ldots, A_k\}$. If Y is an integrable random variable ($\mathcal{F}$-measurable), we compute

$Y' = \mathbf{E}[Y \mid \mathcal{G}]$. Since Y' is $\mathcal{G}$-measurable, we can write $Y' = \alpha_1 \mathbf{1}_{A_1} + \ldots + \alpha_k \mathbf{1}_{A_k}$. Equating its integral over A_i with that of Y, we arrive at $\alpha_i \mathbf{P}(A_i) = \int_{A_i} Y d\mathbf{P}$. Thus,

$$Y' = \sum_{i=1}^{k} \left(\frac{1}{\mathbf{P}(A_i)} \int_{A_i} Y d\mathbf{P} \right) \mathbf{1}_{A_i}.$$

The value of α_i is what you would have seen in basic probability class as the expected value of Y given A_i (just restrict the probability measure to A_i and renormalize by dividing by $\mathbf{P}(A_i)$. Then take expectation of Y w.r.t this new measure).

Example 37

Let X, Y be random variables on $(\Omega, \mathcal{F}, \mathbf{P})$, having a joint density $f(x, y)$ on $\mathbb{R}$. We want to talk of $\mathbf{E}[Y \mid X = x]$. For this, we take $\mathcal{G} = \sigma(X)$, the sigma-algebra generated by X and compute $\mathbf{E}[Y \mid \mathcal{G}]$. What are $\mathcal{G}$-measurable random variables? They are precisely those of the form $\varphi(X)$ for some Borel measurable $\varphi : \mathbb{R} \mapsto \mathbb{R}$ (why?). Let us simply write down the formula and check that it works: $Y' = \varphi(X)$ where

$$\varphi(x) := \begin{cases} \frac{1}{\int_{\mathbb{R}} f(x,y)dy} \int_{\mathbb{R}} y f(x, y) dy & \text{if } \int_{\mathbb{R}} f(x, y) dy > 0 \\ 0 & \text{if } \int_{\mathbb{R}} f(x, y) dy = 0. \end{cases}$$

Clearly Y' is $\mathcal{G}$-measurable (since it is a function of X). Check that $\mathbf{E}[Y' \mathbf{1}_A] = \mathbf{E}[Y \mathbf{1}_A]$ if $A = \{Z \in B\}$ for some $B \in \mathcal{B}_{\mathbb{R}}$. That shows that $Y' = \mathbf{E}[Y \mid \mathcal{G}]$.

It may be confusing for the first time that what we call conditional expectation is a random variable and not a number. But that is indeed the point. First we conceptualize an experiment which tells us for each element of $\mathcal{G}$, whether or not it has occurred. Then depending on the outcome of the experiment, we update our probabilities of event or expectations of random variables. In other words, the update is a function of the outcome of the experiment, hence a random variable.

25. Measure determining classes of random variables

As we have emphasized before, events (when identified with their indicator functions) are a special case of random variables. Thus, often to prove a statement about all integrable random variables, we prove it first for indicators, then for simple functions, then for positive random variables and finally for all integrable random variables.

The other direction can also be useful. To prove a statement about probabilities of events, we generalize the statement to expectations of random variables, prove it for a suitable sub-class of random variables, extend it to all integrable random variables and then specialize to indicators to

get the statement for probabilities of events! The reason this is useful is that there are sub-classes of random variables that are sometimes easier than indicators to work with.

For example, if μ is a Borel probability measure on $\mathbb{R}^n$, the space of continuous functions on $\mathbb{R}^n$, or even smooth functions on $\mathbb{R}^n$ are nice sub-classes of random variables in the following sense.

Proposition 35

The numbers $\mathbf{E}[f(X)]$ as f varies over $C_b(\mathbb{R})$ determine the distribution of X. Equivalently, if $\mu, \nu \in \mathcal{P}(\mathbb{R})$ and $\mathbf{E}_\mu[f] = \mathbf{E}_\nu[f]$ for all $f \in C_b(\mathbb{R})$, then $\mu = \nu$.

Proof. Given any $x \in \mathbb{R}^n$, we can recover $F(x) = \mathbf{E}[\mathbf{1}_{A_x}]$, where $A_x = (-\infty, x_1] \times \ldots \times (-\infty, x_n]$ as follows. For any $\delta > 0$, let $f(y) = \min\{1, \delta^{-1} d(y, A^c_{x+\delta\mathbf{1}})\}$, where d is the L_∞ metric on $\mathbb{R}^n$. Then, $f \in C_b(\mathbb{R})$, $f(y) = 1$ if $y \in A_x$, $f(y) = 0$ if $y \notin A_{x+\delta\mathbf{1}}$ and $0 \leq f \leq 1$. Therefore, $F(x) \leq \mathbf{E}[f \circ X] \leq F(x + \delta\mathbf{1})$. Let $\delta \downarrow 0$, invoke right continuity of F to recover $F(x)$. ■

Much smaller sub-classes of functions are also sufficient to determine the distribution of X.

Lemma 36

Suppose μ, ν are two Borel probability measures on $\mathbb{R}$ such that $\mathbf{E}_\mu[f] = \mathbf{E}_\nu[f]$ for all $f \in C_c^\infty(\mathbb{R})$. Then $\mu = \nu$. Equivalently, the distribution of a random variable X is determined by the numbers $\mathbf{E}[f(X)]$ as f varies over $C_c^\infty(\mathbb{R})$.

Proof. Fix $a < b$. We claim that there exist $f_n \in C_c^\infty(\mathbb{R})$ such that $f_n(x) \uparrow \mathbf{1}_{(a,b)}(x)$ for all x. In particular, $f_n \uparrow \mathbf{1}_{(a,b)}$ $a.s.[\mu]$ and $f_n \uparrow \mathbf{1}_{(a,b)}$ $a.s.[\nu]$ (Caution: If we take the closed interval $[a, b]$, such f_n may not exist). Hence, by MCT, we get $\mathbf{E}_\mu[f_n] \uparrow \mu(a, b)$ and $\mathbf{E}_\nu[f_n] \uparrow \nu(a, b)$. By the hypothesis, $\mathbf{E}_\mu[f_n] = \mathbf{E}_\nu[f_n]$ for all n and hence $\mu(a, b) = \nu(a, b)$. This is true for all $a < b$ and therefore, $\mu = \nu$.

To show the existence of f_n as above, recall that the function

$$g(x) := \begin{cases} Ce^{-1/(1-|x|^2)} & \text{if } |x| \leq 1 \\ 0 & \text{if } |x| > 1 \end{cases}$$

defines a smooth function that vanishes outside $(-1, 1)$. We fix C so that $g(\cdot)$ is a probability density and let G be the corresponding distribution function, i.e., $G(x) = \int_{-\infty}^x g(u)du$. Clearly G is smooth, $G(x) = 0$ for $x < -1$ and $G(x) = 1$ for $x > +1$. Then, $G(n(x - a) - 1)$ vanishes for $x < a$, equals 1 for $x > a + \frac{2}{n}$. Finally, set $f_n(x) = G(n(x - a) - 1)G(n(b - x) - 1)$ and check that f_n satisfies the given properties. ■

26. MEAN, VARIANCE, MOMENTS

Expectations of certain functionals of random variables are important enough to have their own names.

Definition 11

Let X be a r.v. Then, $\mathbf{E}[X]$ (if it exists) is called the *mean* or *expected value* of X. $\text{Var}(X) := \mathbf{E}\left[(X - \mathbf{E}X)^2\right]$ is called the *variance* of X, and its square root is called the *standard deviation* of X. The standard deviation measures the spread in the values of X or one way of measuring the uncertainty in predicting X. Another such measure, not very convenient to use, is the *mean absolute deviation* $\mathbf{E}[|X - \mathbf{E}[X]|]$. For any $p \in \mathbb{N}$, if it exists, $\mathbf{E}[X^p]$ is called the p^{th}*-moment* of X. The function ψ defined as $\psi(\lambda) := \mathbf{E}[e^{\lambda X}]$ is called the *moment generating function* of X. Note that the m.g.f of a non-negative r.v. exists for all $\lambda < 0$. It may or may not exist for some $\lambda > 0$ also. A similar looking object is the *characteristic function* of X, define by $\varphi(\lambda) := \mathbf{E}[e^{i\lambda X}] := \mathbf{E}[\cos(\lambda X)] + i\mathbf{E}[\sin(\lambda X)]$. This exists for all $\lambda \in \mathbb{R}$ since bounded random variables are integrable. All these quantities depend only on the distribution of X and not on the details of the probability space on which X is defined.

For two random variables X, Y on the same probability space, we define their *covariance* to be $\text{Cov}(X, Y) := \mathbf{E}[(X - \mathbf{E}X)(Y - \mathbf{E}Y)] = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]$. The *correlation coefficient* is measured by $\frac{\text{Cov(X,Y)}}{\sqrt{\text{Var(X)Var(Y)}}}$. The correlation coefficient lies in $[-1, 1]$ and measures the association between X and Y. A correlation of 1 implies $X = aY + b$ $a.s.$ for some $a, b \in \mathbb{R}$ with $a > 0$ while a correlation of -1 implies $X = aY + b$ $a.s.$ with $a < 0$. Like with expectation and variance, covariance and correlation depend only on the joint distribution of X and Y.

Exercise 25

(1) Express the mean, variance, moments of $aX + b$ in terms of those for X.

(2) Show that $\text{Var}(X) = \mathbf{E}[X^2] - \mathbf{E}[X]^2$.

(3) Compute mean, variance and moments of the Normal, exponential and other distributions defined in section 12.

Example 38: The exponential distribution

Let $X \sim \text{Exp}(\lambda)$. Then, $\mathbf{E}[X^k] = \int x^k d\mu(x)$ where μ is the p.m on $\mathbb{R}$ with density $\lambda e^{-\lambda x}$ (for $x > 0$). Thus, $\mathbf{E}[X^k] = \int x^k \lambda e^{-\lambda x} dx = \lambda^{-k} k!$. In particular, the mean is λ^{-1}, the variance is $2\lambda^{-2} - (\lambda^{-1})^2 = \lambda^{-2}$.

Example 39: Normal distribution

If $X \sim N(0,1)$, check that the even moments are given by $\mathbf{E}[X^{2k}] = \prod_{j=1}^{k}(2j-1)$.

Remark 13: Moment problem

Given a sequence of numbers $(\alpha_k)_{k\geq 0}$, is there a p.m μ on $\mathbb{R}$ whose k^{th} moment is α_k? If so, is it unique?

This is an extremely interesting question and its solution involves a rich interplay of several aspects of classical analysis (orthogonal polynomials, tridiagonal matrices, functional analysis, spectral theory etc). Note that there are are some non-trivial conditions for (α_k) to be the moment sequence of a probability measure μ. For example, $\alpha_0 = 1$, $\alpha_2 \geq \alpha_1^2$ etc. In the homework you were asked to show that $((\alpha_{i+j}))_{i,j\leq n}$ should be a positive semidefinite matrix for every n. The non-trivial answer is that these conditions are also sufficient!

Note that like proposition 35, the uniqueness question is asking whether $\mathbf{E}[f\circ X]$, as f varies over the space of polynomials, is sufficient to determine the distribution of X. However, uniqueness is not true in general. In other words, one can find two p.m μ and ν on $\mathbb{R}$ which have the same sequence of moments!

27. Product measures and Fubini's theorem

Given two probability spaces $(\Omega_i, \mathcal{F}_i, \mathbf{P}_i)$, $i = 1, 2$, the goal is to define a natural probability measure on the Cartesian product $\Omega_1 \times \Omega_2$. First we decide a natural σ-algebra on the product space and then the measure.

Product σ-algebra: Given two measurable spaces $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$, there are three natural definitions of a σ-algebra on $\Omega = \Omega_1 \times \Omega_2$.

(1) The σ-algebra $\mathcal{R} = \sigma\{A \times B : A \in \mathcal{F}_1,\ B \in \mathcal{F}_2\}$ generated b all "rectangles" (sets of the form $A \times B$).

(2) The σ-algebra $\mathcal{G} = \sigma\{A \times \Omega_2,\ \Omega_1 \times B : A \in \mathcal{F}_1, B \in \mathcal{F}_2\}$ generated by all "cylinder sets" (sets of the form $A \times \Omega_2$ and $\Omega_1 \times B$).

(3) Define the projection maps $\Pi_i : \Omega \to \Omega_i$ by $\Pi_1(x,y) = x$ and $\Pi_2(x,y) = y$. Then define $\mathcal{G}' = \sigma\{\Pi_1, \Pi_2\}$ to be the smallest σ-algebra on Ω for which these projections are measurable.

The first observation is that these definitions give the same σ-algebra, which will be called the *product σ-algebra*. Since $\Pi_1^{-1}(A) = A \times \Omega_2$ for $A \in \mathcal{F}_1$ and $\Pi_2^{-1}(B) = \Omega_1 \times B$ for $B \in \mathcal{F}_2$, it immediately follows that $\mathcal{G} = \mathcal{G}'$. Next, as cylinders are rectangles, clearly $\mathcal{G} \subseteq \mathcal{R}$. But $A \times B =$

$(A \times \Omega_2) \cap (\Omega_1 \times B)$ and hence any rectangle is an intersection of two cylinders. Therefore, $\mathcal{R} \subseteq \mathcal{G}$ and thus $\mathcal{R} = \mathcal{G}$, showing equality of the three sigma algebras. This common sigma algebra is called the product σ-algebra and denoted $\mathcal{F}_1 \otimes \mathcal{F}_2$.

For later purpose, we make some observations.

(1) The set of all rectangles $A \times B$ with $A \in \mathcal{F}_1$ and $B \in \mathcal{F}_2$ forms a π-system. Indeed, $(A_1 \times B_1) \cap (A_2 \times B_2) = (A_1 \cap A_2) \times (B_1 \cap B_2)$.

(2) $(A \times B)^c = (A^c \times \Omega_2) \sqcup (A \times B^c)$. Hence, if $\mathcal{A}$ is the collection of all finite unions of rectangles, then $\mathcal{A}$ is an algebra.

(3) A finite union of rectangles can be written as a finite union of pairwise disjoint rectangles. Thus, $\mathcal{A}$ is also the collection of finite unions of pairwise disjoint rectangles.

For finitely many measurable spaces, $(\Omega_i, \mathcal{F}_i)$, $i \leq n$, it is clear how to define the product sigma algebra on $\Omega_1 \times \ldots \times \Omega_n$. You may take the definition analogous to any of the three definitions given above and check that they agree. Alternately, you may also define it inductively (if $n = 3$, define the product sigma algebra as $(\mathcal{F}_1 \otimes \mathcal{F}_2) \otimes \mathcal{F}_3$) and see that it agrees with the other three definitions (and hence also deduce the associativity property $(\mathcal{F}_1 \otimes \mathcal{F}_2) \otimes \mathcal{F}_3 = \mathcal{F}_1 \otimes (\mathcal{F}_2 \otimes \mathcal{F}_3)$).

Product measure: Let $(\Omega_i, \mathcal{F}_i, \mu_i)$, $1 \leq i \leq n$, be measure spaces. Let $\mathcal{F} = \mathcal{F}_1 \otimes \ldots \otimes \mathcal{F}_n$ be the product sigma algebra on $\Omega := \Omega_1 \times \ldots \times \Omega_n$. A measure μ on $(\Omega, \mathcal{F})$ such that $\mu(A_1 \times \ldots \times A_n) = \prod_{i=1}^n \mu_i(A_i)$ whenever $A_i \in \mathcal{F}_i$ is called a *product measure* and denoted $\mu_1 \otimes \ldots \otimes \mu_n$ (the notation is justified by the theorem below).

Theorem 37: Product measures

Product measure exists and is unique.

Proof. It suffices to take $n = 2$.

The uniqueness part is easy. By the discussion earlier, the collection of all cylinder sets (alternately, rectangles) is a π-system that generates $\mathcal{F}_1 \otimes \mathcal{F}_2$. Since any two product measures agree on rectangles, it follows that they must agree on $\mathcal{F}$. Thus, product measure, if it exists, is unique.

The existence of product measures follows along the lines of the Caratheodary construction using the algebra $\mathcal{A}$ defined earlier. If $A =\in \mathcal{A}$, write $A = R_1 \sqcup \ldots \sqcup R_m$ where $R_j = A_j \times B_j$ are rectangles and define $\mu(A) = \sum_{j=1}^m \mu_1(A_j)\mu_2(B_j)$. Two things need to be checked. (1) The definition is valid (since there may be many ways to write A as a union of pairwise disjoint rectangles). (2) μ is countably additive on the algebra $\mathcal{A}$.

We skip the details of checking[5]. Once that is done, by Caratheodary's theorem, it follows that μ extends to the product sigma algebra. ■

Example 40

If $\Omega_1 = \Omega_2 = \mathbb{R}$ with the Borel sigma algebra on them, then for $\mu_1, \mu_2 \in \mathcal{P}(\mathbb{R})$, the product measure is simply the measure on $\mathbb{R}^2$ with CDF $F(x, y) = F_{\mu_1}(x)F_{\mu_2}(y)$. Indeed, F defined like this is easily checked to be a valid CDF on $\mathbb{R}^2$ and hence corresponds to a measure (but if you see read the proof we gave of that fact, you will see that the proof is almost identical to what is given here - construct the measure on an algebra and then extend it to the sigma algebra - including the details skipped!).

One theorem that we shall state and use is this.

Theorem 38: Fubini's theorem

Let $\mu = \mu_1 \otimes \mu_2$ be a product measure on $\Omega_1 \times \Omega_2$ with the product σ-algebra. If $f : \Omega \to \mathbb{R}_+$ is either a non-negative random variable or an integrable random variable w.r.t μ, then,

(1) For every $x \in \Omega_1$, the function $y \to f(x, y)$ is $\mathcal{F}_2$-measurable and integrable with respect to μ_2 for $a.e.[\mu_1]$ x.

(2) The function $x \to \int f(x, y)d\mu_2(y)$ is $\mathcal{F}_1$-measurable (on the μ_1-measure zero set of x where the integral is not well defined, define the integral to be 0 or in any measurable way).

Further, in both these cases ($f \geq 0$ or $f \in L^1(\mu)$), we have

$$\int_{\Omega} f(z)d\mu(z) = \int_{\Omega_1} \left(\int_{\Omega_2} f(x, y)d\mu_2(y) \right) d\mu_1(x)$$

The same holds with the two co-ordinates interchanged (i.e., you may integrate with respect to μ_1 and then with respect to μ_2).

Proof. Skipped. Attend measure theory class. ■

Here is a simple indication of how one may use this.

Example 41

If $A \in \mathcal{B}_{\mathbb{R}^2}$ has zero Lebesgue measure in $\mathbb{R}^2$, then for $a.e.$ x, the set $A_x = \{y \in \mathbb{R} : (x, y) \in A\}$ has zero Lebesgue measure in $\mathbb{R}$. To see this, consider $\mathbf{1}_A$ and observe

[5] You may consult Dudley's book. We skip details because in the cases that we really need, eg., when $\Omega_i = \mathbb{R}^{d_i}$, we give a different proof later, even for countable products.

that $\int_{\mathbb{R}} \mathbf{1}_A(x,y)d\lambda(y) = \lambda(A_x)$. By Fubini's theorem, $\int_{\mathbb{R}} \lambda(A_x)d\lambda(x) = \lambda_2(A) = 0$. Since $\lambda(A_x) \geq 0$, it follows that $\lambda(A_x) = 0$ for $a.e.\ x$. That was precisely the claim.

Example 42

If X is a non-negative random variable with distribution function F, then $\mathbf{E}[X] = \int_0^\infty (1 - F(t))dt$. To see this, consider $(\Omega, \mathcal{F}, \mathbf{P})$ (on which X is defined) and take its product with $(\mathbb{R}_+, \mathcal{B}, \lambda)$. Let $f(\omega, t) = \mathbf{1}_{X(\omega)>t}$. Check that f is measurable in the product space $\Omega \times \mathbb{R}_+$. Observe that $\int_\Omega f(\omega,t)d\mathbf{P}(\omega) = 1 - F(t)$ while $\int_{\mathbb{R}_+} f(\omega,t)d\lambda(t) = X(\omega)$. Use Fubini's theorem to equate the two iterated integrals $\int_\Omega \int_{\mathbb{R}_+} f(\omega,t)d\lambda(t)d\mathbf{P}(\omega)$ and $\int_{\mathbb{R}_+} \int_\Omega f(\omega,t)d\mathbf{P}(\omega)d\lambda(t)$ to get $\mathbf{E}_\mathbf{P}[X] = \int_{\mathbb{R}_+}(1 - F(t))dt$.

28. Infinite products

Now we want to consider a product of infinitely many probability spaces.

Product σ-algebra: Let I be an arbitrary index set and let $(\Omega_i, \mathcal{F}_i)$, $i \in I$ be measurable spaces. Let $\Omega = \times_{i\in I}\Omega_i$. Again, we have three options for a σ-algebra on Ω.

(1) A rectangle is a set of the form $\times_{i\in I}A_i$ where $A_i \in \mathcal{F}_i$ for each $i \in I$. Let $\mathcal{R}$ be the σ-algebra generated by all rectangles.

(2) A cylinder set is a set of the form $\{\omega \in \Omega : \omega_{i_1} \in A_1, \ldots, \omega_{i_n} \in A_n\}$ for some $n \geq 1$, some $i_1, \ldots, i_n \in I$ and $A_1 \in \mathcal{F}_{i_1}, \ldots, A_n \in \mathcal{F}_{i_n}$. Let $\mathcal{C}$ denote the collection of all cylinder sets and let $\mathcal{G} = \sigma(\mathcal{C})$.

(3) Define the projection maps $\Pi_i : \Omega \to \Omega_i$ by $\Pi_i((x_i)_{i\in I}) = x_i$. Then define $\mathcal{G}' = \sigma\{\Pi_i : i \in I\}$ to be the smallest σ-algebra on Ω for which all these projections are measurable.

Again, $\mathcal{G} = \mathcal{G}'$. Indeed, the cylinder set $\{\omega \in \Omega : \omega_{i_1} \in A_1, \ldots, \omega_{i_n} \in A_n\}$ is precisely $\Pi_{i_1}^{-1}(A_1) \cap \ldots \cap \Pi_{i_n}^{-1}(A_n)$. This shows that cylinders are in $\mathcal{G}'$ and that Π_i are measurable with respect to $\mathcal{G}$. Consequently, $\mathcal{G} = \mathcal{G}'$ and we shall refer to it as the product σ-algebra (or cylinder σ-algebra).

However, $\mathcal{G}$ and $\mathcal{R}$ are not necessarily the same. If I is countable, then the equality is true but not in general if I is uncountable. Let us see why. First of all, cylinders are rectangles and hence $\mathcal{G} \subseteq \mathcal{R}$. It is the other way inclusion that we should worry about.

Suppose I is countable, without loss of generality $I = \mathbb{N}$. Then any rectangle $\times_i A_i$ can be written as the countable intersection $\cap_n B_n$ where $B_n = A_1 \times \ldots \times A_n \times \Omega_{n+1} \times \Omega_{n+2} \ldots$ is a cylinder set. This shows that $\times_i A_i$ is in $\mathcal{G}$ and hence $\mathcal{R} \subseteq \mathcal{G}$. Thus, when I is countable, $\mathcal{R} = \mathcal{G}$. To understand what happens in general, we make the following claim.

Claim 39

Every set in the cylinder σ-algebra is determined by countably many co-ordinates. That is, if $A \in \mathcal{G}$, then there exists a countable set $J \subseteq I$ such that $A \in \sigma\{\Pi_j : j \in J\}$.

Proof. Let $\hat{\mathcal{G}}$ be the collection of all $A \in \mathcal{G}$ that are determined by countably many co-ordinates. If $A \in \sigma\{\Pi_j : j \in J\}$ then $A^c \in \sigma\{\Pi_j : j \in J\}$. Further, if $A_n \in \sigma\{\Pi_j : j \in J_n\}$ for some countable sets $J_n \subseteq I$, then $\cup_n A_n \in \sigma\{\Pi_j : j \in \cup_n J_n\}$. Lastly, $\emptyset \in \hat{\mathcal{G}}$. Thus, $\hat{\mathcal{G}}$ is a σ-algebra. Obviously $\hat{\mathcal{G}}$ contains all cylinder sets and therefore it follows that $\hat{\mathcal{G}} = \mathcal{G}$, proving the claim. ■

As a corollary, if I is uncountable and A_i are proper subsets of Ω_i (possible if Ω_i contain at least two points each!), then the rectangle $\times_{i\in I} A_i$ is not in the cylinder σ-algebra. Thus, whenever Ω_i are not singletons, then the two sigma algebras necessarily differ.

Now that we understand the difference between the two σ-algebras, in the uncountable product, should we consider $\mathcal{R}$ or $\mathcal{G}$? We shall always consider the cylinder σ-algebra $\mathcal{G}$ which will henceforth be denoted $\otimes_{i\in I}\mathcal{F}_i$. We state two reasons. (1) The σ-algebra $\mathcal{R}$ turns out to be too big to support any useful probability measures (just as the power set σ-algebra on $\mathbb{R}$ is too big). (2) In the case when Ω_i are metric spaces (or topological spaces) and $\mathcal{F}_i = \mathcal{B}_{\Omega_i}$, then $\mathcal{G}$ is exactly the Borel σ-algebra on Ω endowed with the product topology. Actually the second reason merely motivates you to brush up the definition of product topology and then you wonder why the product topology was defined that way (why not say that $\times_i A_i$ is open if each A_i is open in Ω_i)? The reason is similar to the first, that is, such a topology is too big to be interesting!

Exercise 26

Show the statement claimed above, that the product σ-algebra on a product of topological spaces is the Borel σ-algebra of the product topology. [**Note:** If you are not familiar with general topological spaces, do this exercise for countable products of metric spaces. Uncountable products of metric spaces are usually not metrizable, hence the suggestion to restrict to countable products.]

Despite all this discussion, we shall consider only countable products in this course. That suffices to cover all cases of interest in probability theory! Recall that in this case, the sigma algebras $\mathcal{R}$ and $\mathcal{G}$ coincide.

Product measure: Let $(\Omega_i, \mathcal{F}_i, \mu_i)$ be probability spaces indexed by $i \in I$. Let $\Omega = \times_{i\in I}\Omega_i$ endowed with the product σ-algebra $\mathcal{F} = \otimes_{i\in I}\mathcal{F}_i$. A probability measure μ on $\mathcal{F}$ is called a product measure of μ_is if for any cylinder set of the form $A = \{\omega \in \Omega : \omega_{i_1} \in A_1, \ldots, \omega_{i_n} \in A_n\}$ we have $\mu(A) = \mu_{i_1}(A_1)\ldots\mu_{i_n}(A_n)$.

Theorem 40: Existence and uniqueness of product measure

For any product of probability spaces, the product measure exists and is unique.

Proof. We can follow the same proof as in the case of finite products. The set of cylinders $\mathcal{C}$ is a π-system and the collection $\mathcal{A}$ of finite unions of pairwise disjoint subsets of $\mathcal{C}$ is an algebra. On $\mathcal{A}$ define the measure in the only natural way, and check that it is well-defined and countably additive (on the algebra). Invoke Caratheodary to conclude that the measure extends to the product sigma algebra. Uniqueness is trivial by the $\pi - \lambda$ theorem (since any two product measures agree on cylinder sets). ■

The reason we have skipped details and given a sketchy proof is that shortly we shall give a different proof in cases of interest. More precisely, we shall take I to be countable, each Ω_i to be $\mathbb{R}^{d_i}$ for some d_i, the sigma algebras to be $\mathcal{B}_{\Omega_i}$ and μ_i to be Borel probability measures. In this situation, we shall show that existence of the product measure $\otimes \mu_i$ by realizing it as the push-forward of Lebesgue measure under a suitable $T : [0,1] \to \Omega = \times_i \Omega_i$. The theorem is as follows.

Theorem 41

Let $\Omega_i = \mathbb{R}^{d_i}$ for $i \in \mathbb{N}$ and let $\mu_i \in \mathcal{P}(\mathbb{R}^{d_i})$ (on the Borel sigma algebra). Then, the product measure $\mu = \otimes_{i \in N} \mu_i$ exists on $\Omega := \times_i \Omega_i$ endowed with the product sigma algebra.

Although the situation described in Theorem 41 covers all cases of actual interest to probabilists, there is some value in the more general theorem Theorem 40. Most importantly, it clarifies that no special properties of $\mathbb{R}^d$ (either as a topological space or any other structure it has) are necessary to construct product measures.

29. Independence

Definition 12: Independence

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space.

- Let $\mathcal{G}_1, \ldots, \mathcal{G}_k$ be sub-sigma algebras of $\mathcal{F}$. We say that $\mathcal{G}_i$ are *independent* if for every $A_1 \in \mathcal{G}_1, \ldots, A_k \in \mathcal{G}_k$, we have $\mathbf{P}(A_1 \cap A_2 \cap \ldots \cap A_k) = \mathbf{P}(A_1) \ldots \mathbf{P}(A_k)$.
- Random variables $X_1, \ldots, X_n$ on $\mathcal{F}$ are said to be independent if $\sigma(X_1), \ldots, \sigma(X_n)$ are independent.
- An arbitrary collection of σ-algebras $\mathcal{G}_i$, $i \in I$, (each $\mathcal{G}_i$ contained in $\mathcal{F}$) are said to be independent if every finite sub-collection of them is independent. Same applies for random variables.

How does this compare with the definitions we have seen in basic probability class?

- Since $\sigma(X) = \{X^{-1}(A) : A \in \mathcal{B}_{\mathbb{R}}\}$ for a real-valued random variable X, the definition above is equivalent to saying that $\mathbf{P}(X_i \in A_i \; i \leq k) = \prod_{i=1}^{k} \mathbf{P}(X_i \in A_i)$ for any $A_i \in \mathcal{B}(\mathbb{R})$. The same definition can be made for random variables X_i taking values in some metric space (Λ_i, d_i), but then A_i must be a Borel subset of Λ_i.
- Events $A_1, \ldots, A_k$ are said to be independent if $\mathbf{1}_{A_1}, \ldots, \mathbf{1}_{A_k}$ are independent. This is equivalent to *either* of the following sets of 2^n conditions:
  (1) $\mathbf{P}(A_{j_1} \cap \ldots \cap A_{j_\ell}) = \mathbf{P}(A_{j_1}) \ldots \mathbf{P}(A_{j_\ell})$ for any $1 \leq j_1 < j_2 < \ldots < j_\ell \leq k$.
  (2) $\mathbf{P}(A_1^{\pm} \cap A_2^{\pm} \cap \ldots \cap A_n^{\pm}) = \prod_{k=1}^{n} \mathbf{P}(A_k^{\pm})$ where we use the notation $A^+ = A$ and $A^- = A^c$.
  The second is clear, since $\sigma(A_k) = \{\emptyset, \Omega, A_k, A_k^c\}$. The equivalence of the first and second is an exercise.

Some remarks are in order.

(1) Independence is defined with respect to a fixed probability measure $\mathbf{P}$.

(2) It would be convenient if we need check the condition in the definition only for a sufficiently large class of sets. However, if $\mathcal{G}_i = \sigma(S_i)$, and for every $A_1 \in S_1, \ldots, A_k \in S_k$ if we have $\mathbf{P}(A_1 \cap A_2 \cap \ldots \cap A_k) = \mathbf{P}(A_1) \ldots \mathbf{P}(A_k)$, we *cannot* conclude that $\mathcal{G}_i$ are independent! If S_i are π-systems, this is indeed true (see below).

(3) Checking pairwise independence is insufficient to guarantee independence. For example, suppose X_1, X_2, X_3 are independent and $\mathbf{P}(X_i = +1) = \mathbf{P}(X_i = -1) = 1/2$. Let $Y_1 = X_2X_3$, $Y_2 = X_1X_3$ and $Y_3 = X_1X_2$. Then, Y_i are pairwise independent but not independent.

Lemma 42

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. Assume that $\mathcal{G}_i = \sigma(S_i) \subseteq \mathcal{F}$, that S_i is a π-system and that $\Omega \in S_i$ for each $i \leq k$. If for every $A_1 \in S_1, \ldots, A_k \in S_k$ if we have $\mathbf{P}(A_1 \cap A_2 \cap \ldots \cap A_k) = \mathbf{P}(A_1) \ldots \mathbf{P}(A_k)$, then $\mathcal{G}_i$ are independent.

Proof. Fix $A_2 \in S_2, \ldots, A_k \in S_k$ and set $\mathcal{F}_1 := \{B \in \mathcal{G}_1 : \mathbf{P}(B \cap A_2 \cap \ldots \cap A_k) = \mathbf{P}(B)\mathbf{P}(A_2) \ldots \mathbf{P}(A_k)\}$. Then $\mathcal{F}_1 \supseteq S_1$ by assumption. We claim that $\mathcal{F}_1$ is a λ-system. Assuming that, by the π-λ theorem, it follows that $\mathcal{F}_1 = \mathcal{G}_1$ and we get the assumptions of the lemma for $\mathcal{G}_1, S_2, \ldots, S_k$. Repeating the argument for S_2, S_3 etc., we get independence of $\mathcal{G}_1, \ldots, \mathcal{G}_k$.

To prove that $\mathcal{F}_1$ is a λ system is straightforward. If $B_n \uparrow B$ and $B_n \in \mathcal{F}_1$, then $B \in \mathcal{F}$ and $\mathbf{P}(B_n \cap A_2 \cap \ldots \cap A_k) \uparrow \mathbf{P}(B \cap A_2 \cap \ldots \cap A_k)$ and $\mathbf{P}(B_n) \prod_{j=2}^{k} \mathbf{P}(A_j) \uparrow \mathbf{P}(B) \prod_{j=2}^{k} \mathbf{P}(A_j)$. Hence $B \in \mathcal{F}_1$. Similarly, check that if $B_1 \subseteq B_2$ and both are in $\mathcal{F}_1$, then $B_2 \setminus B_1 \in \mathcal{F}_1$. Lastly, $\Omega \in S_1 \subseteq \mathcal{F}_1$ by assumption. Thus, $\mathcal{F}_1$ is a λ-system. ■

Remark 14

If $A_1, \ldots, A_k$ are events, then $\mathcal{G}_i = \{\emptyset, A_i, A_i^c, \Omega\}$ is generated by the π-system $S_i = \{A_i\}$. However, checking the independence condition for the generating set (which is just one equation $\mathbf{P}(A_1 \cap \ldots \cap A_k) = \prod_{j=1}^k \mathbf{P}(A_j)$) does not imply independence of $A_1, \ldots, A_k$. This shows that the condition that S_i should contain Ω is not redundant in the above Lemma!

Corollary 43

(1) Random variables $X_1, \ldots, X_k$ are independent if and only if for every $t_1, \ldots t_k \in \mathbb{R}$ we have $\mathbf{P}(X_1 \leq t_1, \ldots, X_k \leq t_k) = \prod_{j=1}^k \mathbf{P}(X_j \leq t_j)$.

(2) Suppose $\mathcal{G}_\alpha$, $\alpha \in I$ are independent. Let $I_1, \ldots, I_k$ be pairwise disjoint subsets of I. Then, the σ-algebras $\mathcal{F}_j = \sigma\left(\cup_{\alpha \in I_j} \mathcal{G}_\alpha\right)$ are independent.

(3) If $X_{i,j}$, $i \leq n$, $j \leq n_i$, are independent, then for any Borel measurable $f_i : \mathbb{R}^{n_i} \to \mathbb{R}$, the r.v.s $f_i(X_{i,1}, \ldots, X_{i,n_i})$ are also independent.

Proof. (1) The sets $(-\infty, t]$ form a π-system that generates $\mathcal{B}(\mathbb{R})$ and hence $S_i := \{X_i^{-1}(-\infty, t] : t \in \mathbb{R}\}$ is a π-system that generates $\sigma(X_i)$.

(2) For $j \leq k$, let S_j be the collection of finite intersections of sets A_i, $i \in I_j$. Then S_j are π-systems and $\sigma(S_j) = \mathcal{F}_j$.

(3) Infer (3) from (2) by considering $\mathcal{G}_{i,j} := \sigma(X_{i,j})$ and observing that $f_i(X_{i,1}, \ldots, X_{i,k}) \in \sigma(\mathcal{G}_{i,1} \cup \ldots \cup \mathcal{G}_{i,n_i})$. ■

So far, we stated conditions for independence in terms of probabilities of events. As usual, they generalize to conditions in terms of expectations of random variables.

Lemma 44

(1) Sigma algebras $\mathcal{G}_1, \ldots, \mathcal{G}_k$ are independent if and only if for every $\mathcal{G}_i$-measurable, bounded random variable X_i, for $1 \leq i \leq k$, we have $\mathbf{E}[X_1 \ldots X_k] = \prod_{i=1}^k \mathbf{E}[X_i]$.

(2) In particular, random variables $Z_1, \ldots, Z_k$ (Z_i is an n_i dimensional random vector) are independent if and only if $\mathbf{E}[\prod_{i=1}^k f_i(Z_i)] = \prod_{i=1}^k \mathbf{E}[f_i(Z_i)]$ for any bounded Borel measurable functions $f_i : \mathbb{R}^{n_i} \to \mathbb{R}$.

We say 'bounded measurable' just to ensure that expectations exist. The proof goes inductively by fixing $X_2, \ldots, X_k$ and then letting X_1 be a simple r.v., a non-negative r.v. and a general bounded measurable r.v.

Proof. (1) Suppose $\mathcal{G}_i$ are independent. If X_i are $\mathcal{G}_i$ measurable then it is clear that X_i are independent and hence $\mathbf{P}(X_1, \ldots, X_k)^{-1} = \mathbf{P}X_1^{-1} \otimes \ldots \otimes \mathbf{P}X_k^{-1}$. Denote $\mu_i := \mathbf{P}X_i^{-1}$ and apply Fubini's theorem (and change of variables) to get

$$\begin{aligned}\mathbf{E}[X_1 \ldots X_k] &\overset{\text{c.o.v}}{=} \int_{\mathbb{R}^k} \prod_{i=1}^k x_i d(\mu_1 \otimes \ldots \otimes \mu_k)(x_1, \ldots, x_k) \\ &\overset{\text{Fub}}{=} \int_{\mathbb{R}} \ldots \int_{\mathbb{R}} \prod_{i=1}^k x_i d\mu_1(x_1) \ldots d\mu_k(x_k) \\ &= \prod_{i=1}^k \int_{\mathbb{R}} u d\mu_i(u) \overset{\text{c.o.v}}{=} \prod_{i=1}^k \mathbf{E}[X_i].\end{aligned}$$

Conversely, if $\mathbf{E}[X_1 \ldots X_k] = \prod_{i=1}^k \mathbf{E}[X_i]$ for all $\mathcal{G}_i$-measurable functions X_is, then applying to indicators of events $A_i \in \mathcal{G}_i$ we see the independence of the σ-algebras $\mathcal{G}_i$.

(2) The second claim follows from the first by setting $\mathcal{G}_i := \sigma(Z_i)$ and observing that a random variable X_i is $\sigma(Z_i)$-measurable if and only if (see remark following the proof) $X = f \circ Z_i$ for some Borel measurable $f : \mathbb{R}^{n_i} \to \mathbb{R}$. ■

Remark 15

We stated a fact that if X is a real-valued random variable and $Y \in \sigma(X)$, then $Y = f(X)$ for some $f : \mathbb{R} \to \mathbb{R}$ that is Borel measurable. Why is that so?

If $X(\omega) = X(\omega')$, then it is clear that any set $A \in \sigma(X)$ either contains both ω, ω' or excludes both (this was an exercise). Consequently, we must have $Y(\omega) = Y(\omega')$ (otherwise, if $Y(\omega) < a < Y(\omega')$ for some $a \in \mathbb{R}$, then the set $Y < a$ could not be in $\sigma(X)$, as it contains ω but not ω'). This shows that $Y = f(X)$ for some function $f : \mathbb{R} \to \mathbb{R}$. But why is f measurable? Indeed, one should worry a little, because the correct statement is not that f is measurable, but that f may be chosen to be measurable. For example, if X is the constant 0 and Y is the constant 1, then all we know is $f(0) = 1$. We shall have $Y = f(X)$ however we define f on $\mathbb{R} \setminus \{0\}$ (in particular, we may make f non-measurable!).

One way out is to use the fact that the claim is true for simple random variables and that every random variable can be written as a pointwise limit of simple random variables (see exercise below). Consequently, $Y = \lim Y_n$, where Y_n is a $\sigma(X)$-measurable simple random variable and hence $Y_n = f_n(X)$ for some Borel measurable $f_n : \mathbb{R} \to \mathbb{R}$. Let $f = \limsup f_n$, also Borel measurable. But $Y = f(X)$.

Exercise 27

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. Then every random variable on Ω is a pointwise limit of simple random variables.

30. Independent sequences of random variables

First we make the observation that product measures and independence are closely related concepts. Indeed, if $X_1, \ldots, X_k$ are random variables on a common probability space, then the following statements are equivalent.

(1) $X_1, \ldots, X_n$ are independent.

(2) If $X = (X_1, \ldots, X_n)$, then $\mathbf{P} \circ X^{-1}$ is the product measure $\mathbf{P}X_1^{-1} \otimes \ldots \otimes \mathbf{P}X_k^{-1}$.

To see this, use the definition of independence and of product measure. The same holds for infinite collections of random variables too. That is, if X_i, $i \in I$ are random variables on a common probability space, then they are independent if and only if $\mathbf{P} \circ X^{-1} = \otimes_{i \in I} \mathbf{P} \circ X_i^{-1}$, where $X : \Omega \to \mathbb{R}^I$ is defined as $[X(\omega)](i) = X_i(\omega)$. Of course, the sigma-algebra on $\mathbb{R}^I$ is the product of Borel sigma algebras on the real line.

Theorem 40 asserts the existence of the product probability measure on the product of any given collection of probability spaces. We sketched the proof, which is via Caratheodary's method of constructing a measure on the algebra of cylinder sets and then extending it to the product sigma algebra. We skipped checking that the measure defined on the algebra was countably additive, a key point in the construction.

In this section, we restrict to countable products of $(\mathbb{R}, \mathcal{B}_{\mathbb{R}}, \mu_i)$ and show the existence of the product measure in a different way. This proof easily extends to the product of $(\mathbb{R}^{d_i}, \mathcal{B}_{\mathbb{R}^{d_i}}, \mu_i)$ or even of $(\Omega_i, \mathcal{F}_i, \mu_i)$ provided each μ_i is the push-forward of λ (Lebesgue measure on $[0,1]$). However, we shall do this in the language of random variables rather than measures, something one must get used to in probability theory. To do that, we observe that the following questions are equivalent.

(1) **Question 1:** Given $\mu_i \in \mathcal{P}(\mathbb{R})$, $i \geq 1$, does there exist a probability space with independent random variables X_i having distributions μ_i?

(2) **Question 2:** Given $\mu_i \in \mathcal{P}(\mathbb{R})$, $i \geq 1$, does there exist a p.m μ on $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty))$ such that $\mu(A_1 \times \ldots \times A_n \times \mathbb{R} \times \mathbb{R} \times \ldots) = \prod_{i=1}^n \mu_i(A_i)$? In other words, does the product measure exist?

The equivalence is easy to see. Suppose we answer the first question by finding an $(\Omega, \mathcal{F}, \mathbf{P})$ with *independent* random variables $X_i : \Omega \to \mathbb{R}$ such that $X_i \sim \mu_i$ for all i. Then, $X : \Omega \to \mathbb{R}^\infty$ defined by $X(\omega) = (X_1(\omega), X_2(\omega), \ldots)$ is measurable w.r.t the relevant σ-algebras (why?). Then,

let $\mu := \mathbf{P}X^{-1}$ be the pushforward p.m on $\mathbb{R}^\infty$. Clearly

$$\begin{aligned}\mu(A_1 \times \ldots \times A_n \times \mathbb{R} \times \mathbb{R} \times \ldots) &= \mathbf{P}\left(X_1 \in A_1, \ldots, X_n \in A_n\right) \\ &= \prod_{i=1}^n \mathbf{P}(X_i \in A_i) = \prod_{i=1}^n \mu_i(A_i).\end{aligned}$$

Thus μ is the product measure required by the second question.

Conversely, if we could construct the product measure on $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty))$, then we could take $\Omega = \mathbb{R}^\infty$, $\mathcal{F} = \mathcal{B}(\mathbb{R}^\infty)$ and $X_i = \Pi_i$, the i^{th} co-ordinate random variable. Then you may check that they satisfy the requirements of the first question.

The two questions are thus equivalent, but what is the answer?! It is 'yes', of course or we would not make heavy weather about it.

Proposition 45: [Daniell, Kolmogorov]

Let $\mu_i \in \mathcal{P}(\mathbb{R})$, $i \geq 1$, be Borel p.m on $\mathbb{R}$. Then, there exist a probability space with *independent* random variables $X_1, X_2, \ldots$ such that $X_i \sim \mu_i$.

Proof. We arrive at the construction in three stages.

(1) **Independent Bernoullis:** On the probability space $((0,1), \mathcal{B}, \lambda)$, consider the random variables $X_k : (0,1) \to \mathbb{R}$, where $X_k(\omega)$ is defined to be the k^{th} digit in the binary expansion of ω (see Section 11 for convention regarding binary expansion). Then by an earlier homework exercise, $X_1, X_2, \ldots$ are independent Bernoulli$(1/2)$ random variables.

(2) **Independent uniforms:** Note that as a consequence[6], on any probability space, if Y_i are i.i.d. Ber$(1/2)$ variables, then $U := \sum_{n=1}^\infty 2^{-n} Y_n$ has uniform distribution on $[0,1]$. Consider again the canonical probability space and the r.v. X_i, and set $U_1 := X_1/2 + X_3/2^2 + X_5/2^3 + \ldots$, $U_2 := X_2/2 + X_6/2^2 + \ldots$, $U_3 = X_4/2 + X_{12}/2^2 + \ldots$ etc. (in short, let $g : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ be an injection and define $Y_k = \sum_{j=1}^\infty X_{g(k,j)} 2^{-j}$). Clearly, U_i are i.i.d. Unif$[0,1]$.

(3) **Arbitrary distributions:** For a p.m. μ, recall the left-continuous inverse G_μ that had the property that $G_\mu(U) \sim \mu$ if $U \sim U[0,1]$. Suppose we are given p.m.s $\mu_1, \mu_2, \ldots$. On the canonical probability space, let U_i be i.i.d uniforms constructed as before. Define $X_i :=$

[6]Let us be pedantic and show this: Suppose Y_i are independent Bernoullis on $(\Omega, \mathcal{F}, \mathbf{P})$ and $T = (Y_1, Y_2, \ldots) : \Omega \to \{0,1\}^\infty$. Then $\mu := \mathbf{P} \circ T^{-1}$ is the product Bernoulli measure on $\{0,1\}^\infty$. Let $V : \{0,1\}^\infty \to \mathbb{R}$ be defined as $V(x) = \sum_k x_k 2^{-k}$ so that $(V \circ T)(\omega)$ is precisely $\sum_k Y_k(\omega) 2^{-k}$, the random variable that we want. By the reasoning in Lemma 31, we see that $\mathbf{P} \circ (V \circ T)^{-1} = \mu \circ V^{-1}$. This shows that the distribution of $\sum_k Y_k 2^{-k}$ does not depend on the original probability space. But for X_k as before, we get $\sum_k X_k 2^{-k}$ has uniform$([0,1])$ distribution, hence the same holds on any probability space. Again, we emphasize the unimportance of the original probability space, what matters is the joint distribution of the random variables that we are interested in.

$G_{\mu_i}(U_i)$. Then, X_i are independent and $X_i \sim \mu_i$. Thus we have constructed an independent sequence of random variables having the specified distributions. ■

This proof does not work for uncountable products. However, it does work for a countable product of $(\Omega_i, \mathcal{F}_i, \mu_i)$, provided each μ_i is a pushforward of Lebesgue measure, that is, $\mu_i = \mathbf{P} \circ T_i^{-1}$ for some $T_i : [0,1] \to \Omega_i$. The only change needed is to set $X_i = T_i(U_i)$ (instead of $G_{\mu_i}(U_i)$) in the last step. As we know, all Borel probability measures on $\mathbb{R}^d$ are push-forwards of Lebesgue measure and hence, the above proof works if $\Omega_i = \mathbb{R}^{d_i}$ and $\mu_i \in \mathcal{P}(\mathbb{R}^{d_i})$. The following exercise (not trivial!) shows that it is not possible to get uncountable products in this way.

Exercise 28

Show that there do not exist uncountably many independent, non-constant random variables on $([0,1], \mathcal{B}, \lambda)$. Deduce that the measure $\otimes_{x\in\mathbb{R}}\text{Ber}(1/2)$ on $\{0,1\}^{\mathbb{R}}$ with the product sigma-algebra, cannot be realized as the push-forward of Lebesgue measure.

31. Kolmogorov's consistency theorem

A generalization of the theorem on the existence of product measures is to go beyond independence. To motivate it, consider the following question. Given three Borel probability measures μ_i, $i \leq 3$, does there exist a probability space and three random variables X_i such that $X_i \sim \mu_i$? The answer is trivially yes, for example we can take three independent random variables having the distribution μ_i. Alternately, we may take one uniform random variable and set $X_i = G_{\mu_i}(U)$ (then X_i won't be independent).

Having disposed the easy question, what if we specify three Borel probability measures ν_i on $\mathbb{R}^2$ and want $(X_1, X_2) \sim \nu_1$, $(X_2, X_3) \sim \nu_2$ and $(X_1, X_3) \sim \nu_3$? Is it possible to find such random variables? If the first marginal of ν_1 and the first marginal of ν_3 do not agree, then it is not possible (because then we have two distinct specifications for the distribution of X_1!). This is because our specifications were internally inconsistent. The following theorem of Kolmogorov asserts that this is the only obstacle in constructing random variables with specified finite dimensional distributions.

Theorem 46: Kolmogorov's consistency theorem

Let $\Omega_i = \mathbb{R}^{d_i}$ for some $d_i \geq 1$. For each $n \geq 1$ and each $1 \leq i_1 < i_2 < \ldots < i_n$, let $\mu_{i_1,\ldots,i_n}$ be a Borel p.m on $\Omega_{i_1} \times \ldots \times \Omega_{i_n}$. Then the following are equivalent.

(1) There exists a unique Borel probability measure μ on $\times_i \Omega_i$ such that $\mu \circ \Pi_{i_1,\ldots,i_n}^{-1} = \mu_{i_1,\ldots,i_n}$ for any $i_1 < i_2 < \ldots < i_n$ and any $n \geq 1$.

(2) The given family of probability measures satisfy the consistency condition

$$\mu_{i_1,\ldots,i_n}(B \times \Omega_{i_n}) = \mu_{i_1,\ldots,i_{n-1}}(B)$$

for any $B \in \mathcal{B}(\Omega_{i_1} \times \ldots \times \Omega_{i_{n-1}})$ and for any $n \geq 1$ and any $i_1 < i_2 < \ldots < i_n$.

We have stated the consistency theorem for Ω_i that are Euclidean spaces. It can be generalized, but some metric structure on Ω_is is needed. This is in contrast to the situation of product measures, which exist even if Ω_i have no structure.

Alternate form of the consistency condition: Suppose for each $n \geq 1$, we have a probability measure ν_n on $\Omega_1 \times \ldots \times \Omega_n$. Assume that $\nu_{n+1}(A_1 \times \ldots \times A_n \times \Omega_{n+1}) = \nu_n(A_1 \times \ldots \times A_n)$ for all $n \geq 1$ and all $A_i \in \mathcal{F}_i$. Then, for any $1 \leq i_1 < \ldots < i_k$ and any $n \geq i_k$, the probability measure $\nu_n \circ \Pi^{-1}_{i_1,\ldots,i_k}$ on $\Omega_{i_1} \times \ldots \times \Omega_{i_k}$ is the same. If we define this to be $\mu_{i_1,\ldots,i_k}$, then we get a consistent family of probability measures as required in the theorem.

The importance of the consistency theorem comes from having to construct dependent random variables such as Markov chains with given transition probabilities. It also serves as a starting point for even more subtle questions such as constructing stochastic processes such as Brownian motion.

Proof of the consistency theorem. The necessity of the consistency conditions is clear. It is the other way implication that needs proof. ■

32. Applications of the consistency theorem

32.1. **Markov chains.** Consider $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and let $\mu_0 \in \mathcal{P}(\mathbb{R}^d)$ and let $\kappa : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \mapsto \mathbb{R}_+$ be a transition kernel. This means that $y \mapsto \kappa(x, \cdot)$ is a Borel probability measure function for each $x \in \mathbb{R}^d$ and $x \mapsto \kappa(x, A)$ is Borel measurable for each $A \in \mathcal{B}(\mathbb{R}^d)$. Then, define for each $n \geq 1$, a probability measure on $(\mathbb{R}^d)^n$ by

$$\nu_n(A_0 \times A_1 \times \ldots \times A_{n-1}) = \int_{A_0}\int_{A_1}\ldots\int_{A_{n-1}} \kappa(x_{n-2}, dx_{n-1})\kappa(x_{n-3}, dx_{n-2}) \ldots \kappa(x_0, dx_1)d\mu(x_0).$$

for any $A_i \in \mathcal{B}(\mathbb{R}^d)$. It may be easier to parse this expression if we assume that all the measures μ_0 and $\kappa(x, \cdot)$ are absolutely continuous to one measure θ. In this case, write $d\mu_0(x) = \rho(x)d\theta(x)$ and $\kappa(x, dy) = p(x, y)d\theta(y)$ and then

$$\begin{aligned}&\nu_n(A_0 \times A_1 \times \ldots \times A_{n-1})\\ &= \int_{A_0}\int_{A_1}\ldots\int_{A_{n-1}} p(x_{n-2}, x_{n-1})p(x_{n-3}, x_{n-2}) \ldots p(x_0, x_1)\rho(x_0)\, d\theta(x_{n-1}) \ldots d\theta(x_0).\end{aligned}$$

That is, ν_n has density $\rho(x_0)p(x_0,x_1)\dots p(x_{n-2},x_{n-1})$ with respect to $\theta^{\otimes n}$.

It is easy to check that ν_n defines a probability measure on $(\mathbb{R}^d)^n$ and also that $\nu_{n+1}(A_0 \times \ldots \times A_{n-1} \times \mathbb{R}^d) = \nu_n(A_0 \times \ldots \times A_{n-1})$. Consequently, by the alternate form of the consistency condition stated above, we see that there is a probability measure μ on $(\mathbb{R}^d)^{\mathbb{N}}$ (endowed with the Borel/cylinder sigma algebra) such that $\mu \circ \Pi^{-1}_{0,1,\ldots,n-1} = \nu_n$. This measure μ on $\mathbb{R}^{\mathbb{N}}$ is what is called a *Markov chain* with state space $\mathbb{R}^d$, transition kernel p and initial distribution μ_0.

32.2. **Gaussian processes.** Suppose $m : \mathbb{Z} \to \mathbb{R}$ and $\sigma : \mathbb{Z}\times\mathbb{Z} \to \mathbb{R}$. A Gaussian process with mean $\mu(\cdot)$ and covariance $\sigma = (\sigma_{i,j})_{i,j\in\mathbb{Z}}$ is a collection of jointly Gaussian random variables $(X_n)_{n\in\mathbb{Z}}$ such that $\mathbf{E}[X_n] = \mu(n)$ and $\mathrm{Cov}(X_n, X_m) = \sigma(m,n)$.

Question: Does it exist?

First let us note some necessary conditions. If we could construct a Gaussian process Y with mean 0 and we set $X = m + Y$ (i.e., $X_n = m(n) + Y_n$) has mean $m(\cdot)$ and the same covariance as Y. Hence the mean poses no challenge and we assume that it is zero henceforth.

The covariance is more subtle. For example, $\sigma(n,n) = \mathbf{E}[X_n^2]$ cannot be negative. More generally, for any $n \geq 0$ and $i_1 < \ldots < i_n$ and any $c_1, \ldots, c_n \in \mathbb{R}$, we must have

$$0 \leq \mathbf{E}[(c_1X_{i_1} + \ldots + c_nX_{i_n})^2] \;=\; \sum_{p,q=1}^{n} c_pc_q\mathbf{E}[X_{i_p}X_{i_q}] \;=\; \sum_{p,q=1}^{n} c_pc_q\sigma(i_p,i_q).$$

Thus, every principal finite sub-matrix of σ must be positive semi-definite. We now claim that this is also sufficient.

Assume that σ is positive definite in the above sense. Then for any $n \geq 1$ and any $i_1 < \ldots < i_n$, the measure $\mu_{i_1,\ldots,i_n} = N_n(0, (\sigma(i_p,i_q))_{p,q,\leq n})$ is well-defined. This is because positive definiteness allows us to write

$$(\sigma(i_p,i_q))_{p,q\leq n} = BB^t$$

for a $n \times n$ matrix B. Taking $Z_1, \ldots, Z_n$ i.i.d. $N(0,1)$, the distribution of the random vector BZ, where $Z = (Z_1, \ldots, Z_n)^t$ is the desired Gaussian distribution.

From basic properties of Gaussian distributions (marginals of Gaussians are Gaussian) it follows that the family of distributions $\{\mu_{i_1,\ldots,i_n}\}$ is consistent. Hence by the consistency theorem, the Gaussian process with covariance σ exists.

32.3. **Did we really need the consistency theorem?** Actually no! We could have constructed Markov chains and Gaussian processes from the simpler fact that i.i.d. uniform random variables $V_0, V_1, V_2, \ldots$ exist. For Markov chains, to take the kth step, we can use V_k to generate a random variable from the required step distribution (depending on the current location). For Gaussian

process, one can first convert V_k to $Z_k \sim N(0,1)$. Then the Gaussian process can be generated in the form $X = BZ$, where $Z = (Z_1, Z_2, \ldots)^t$ and B is an infinite matrix such that $BB^t = \sigma$ (here the indexing set is $\mathbb{N}$ instead of $\mathbb{Z}$ which of course makes no difference, and B can even be taken to be lower triangular, which avoids infinite sums in computing BB^t).

In fact, every situation of interest to probabilists can be generated from a sequence of independent random variables, and hence on the probability space $([0,1], \mathcal{B}, \lambda)$. The idea is that we construct $X_{n+1} = f_n(U, X_1, \ldots, X_n)$ where f_n is the inverse of the cumulative distribution function of the conditional distribution of X_{n+1} given $\sigma\{X_1, \ldots, X_n\}$. We have not yet defined what conditional distribution means, but in the situations where you know what it means, it should be clear that the above procedure works.