2 Measure Theory and Integration (Lectures 3 & 4)

Before we delve into Markov Chains we need to cover some basic tools to describe and analyse them. Measure Theory is a fundamental building block of modern probability and integration theory. Without it, the theory of probability is rather dull. You are already familiar with continuous and discrete random variables, but these are rather restrictive. For example, consider an experiment in which we flip a coin. If it turns heads, we set $X = 0$, if it turns tails we sample $X \sim \mathcal{U}[0, 1]$. The random variable X is not continuous, since the probability at a single point is non-zero, and it is not discrete as it can attain any value in $[0, 1]$. What kind of a random variable is X? Does it have an expectation? Our intuition tells us we should on average expect a value of $1/4$, but we cannot apply any of the formulas below:

$$\mathbb{E}[X] = \int_\Omega xp(x)\mathrm{d}x \tag{26}$$

$$\mathbb{E}[X] = \sum_{x\in\Omega} xp(x) \tag{27}$$

Random variables like X arise very naturally and will be a basic building block of the Metropolis-Hastings algorithm. Measure Theory equips us with a rich language to talk about them.

Measure Theory studies the notion of a measure — e.g. area, length, mass etc. Conventionally, these quantities are defined by an integral

$$I = \int_\Omega f(x)\mathrm{d}x \tag{28}$$

where, if $f(x)$ represents density, then I represents the total mass of Ω, and if $f(x)$ is represents the height of a surface, then I represents the total volume of the set Ω below the surface. The usual definition of an integral is due to Riemann, which we will briefly review below. The modern definition of an integral is due to Lebesgue, who was the father of Measure Theory. Instead of defining measures through an integral, Lebesgue first defined what a measure is and only then used it to define the integration operation. The two definitions are not contradictory, and for many functions agree. However, as we will see, measure theoretic integration will prove to be much more convenient and general.

The content in these two lectures might be difficult to grasp at first and we recommend reading the material slowly, returning to it as you progress through the course.

2.1 Riemann Integration

The definition of a Riemann integral is based on approximations using step functions.

Definition 4. *We say that $s : [a, b] \to \mathbb{R}$ is a step function, if for some partition $a = t_0 < t_1 < \cdots < t_N = b$:*

$$s(x) = \sum_{i=1}^{N} c_i \mathbb{1}_{[t_{i-1},t_i)}(x) \tag{29}$$

where $c_i \in \mathbb{R}$ are constants.

We define the integral of a step function as

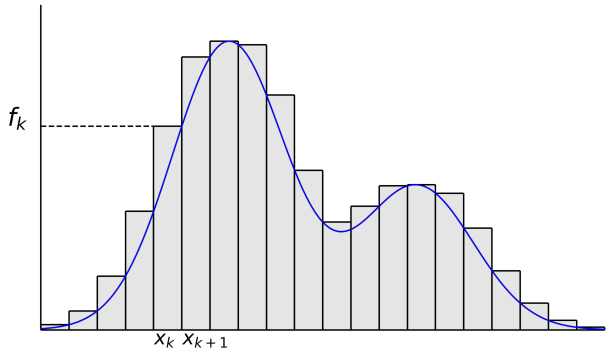


Figure 4: Riemann integral

$$\int_a^b s(x)\mathrm{d}x = \sum_{i=1}^{N} c_i(t_i - t_{i-1}) \tag{30}$$

Given a general function $f : [a, b] \to \mathbb{R}$ we will define its integral by bounding it from above and below by step functions:

Definition 5. *Let $f : [a, b] \to \mathbb{R}$ be a function. Define*

$$I_- = \sup_{s \le f} \int_a^b s(x)\mathrm{d}x \tag{31}$$

$$I_+ = \inf_{s \ge f} \int_a^b s(x)\mathrm{d}x \tag{32}$$

i.e. the "highest" integral of all step functions bounding f from below, and the "lowest" integral of all step functions bounding f from above. If they are equal, i.e. $I = I_- = I_+$, then we say that f is integrable and its integral is defined as

$$\int_a^b f(x)\mathrm{d}x = I \tag{33}$$

There are a few fundamental problems with the above classical definition of an integral. Firstly, it is dependent on the notion of length of the interval (t_{i-1}, t_i). Extending this definition to functions on $\mathbb{R}^d$ should not be too complex — we could define step functions being constant on hypercubes $[a_1, b_1) \times [a_2, b_2) \cdots \times [a_n, b_n)$ with their volume $(b_1 - a_1) \times (b_2 - a_2) \times \cdots \times (b_n - a_n)$ appearing in the definition instead of $(t_i - t_{i-1})$. However, if we wanted to integrate over infinite dimensional spaces (e.g. the space of sequences) this would not work — an infinite dimensional hypercube with an edge of length 2 would have volume $(b_1 - a_1) \times (b_2 - a_2) \times \cdots = \infty$. That is, bounded sets have infinite volume. Putting real numbers aside, what if we want to integrate over a set of functions? How do we define an interval?

Secondly, the Riemann integral is very restrictive regarding the exchange of integrals and limits. That is, often we will have a sequence of approximating functions $f_1, f_2, \ldots$ to the function $f(x) = \lim f_n(x)$ and wonder if the approximation is good enough for us to expect that the integrals of $f_1, f_2, \ldots$ will converge to the integral of f, i.e.

$$\lim_{n\to\infty}\int_a^b f_n(x)\mathrm{d}x = \int_a^b \lim_{n\to\infty} f_n(x)\mathrm{d}x = \int_a^b f(x)\mathrm{d}x \tag{34}$$

To answer this, we need to define what "approximating" means, of which there are many definitions.

Definition 6. *We say that a sequence of functions $f_1, f_2, \ldots$ converges pointwise to a function f if for all x, $\lim_{n\to\infty} f_n(x) = f(x)$. Or, in ϵ-language: if for all x and $\epsilon > 0$, there is a natural number $N = N(\epsilon, x)$ such that $n \geq N \implies |f(x) - f_n(x)| < \epsilon$.*

We write $\lim_{n\to\infty} f_n = f$.

Definition 7. *We say that a sequence of functions $f_1, f_2, \ldots$ converges uniformly to a function f, if the maximum error tends to 0, i.e. $\sup_x |f(x) - f_n(x)| \to 0$. Or, in ϵ-language: if for all $\epsilon > 0$ there is a natural number $N = N(\epsilon)$ such that $n \geq N \implies |f(x) - f_n(x)| < \epsilon$ for all x.*

We then write $f_n \xrightarrow{u} f$.

It should be clear that uniform convergence implies pointwise convergence, but not the other way around. It's also easy to check that exchanging limits and integration under uniform convergence is permitted:

$$\left|\lim_{n\to\infty}\int_a^b f_n(x)\mathrm{d}x - \int_a^b f(x)\mathrm{d}x\right| = \left|\lim_{n\to\infty}\int_a^b f_n(x) - f(x)\mathrm{d}x\right| \tag{35}$$

$$\leq \lim_{n\to\infty}\int_a^b |f_n(x) - f(x)|\mathrm{d}x \tag{36}$$

$$\leq \lim_{n\to\infty}(b-a)\sup_x |f_n(x) - f(x)| = 0 \tag{37}$$

However, this is not true for pointwise convergence, and good examples of this are given in the notebooks and lectures.

2.2 σ-algebras and measures

In this subsection we will define the abstract notion of a measure on an arbitrary set Ω. It can be the set $\mathbb{R}$ or any other set of objects. While reading, it might be helpful to imagine the case $\Omega = \mathbb{R}^2$, where our goal is to define what the notion of an area is. It is tempting to think that a measure should be a function taking points of Ω as an input. Note, however, that by the classical notion of an area any single point on $\mathbb{R}^2$ has an area of 0. A non-zero area is only achieved by a collection of points. Therefore, at least in the case when $\Omega = \mathbb{R}^2$, a measure should be a function of subsets of Ω.

For example, in the contexts of physics, given a material density $f(x)$ and a set $A \subset \Omega$, by integrals such as

$$\int_A f(x)\mathrm{d}x \tag{38}$$

we usually mean the total mass of the set A. Given a function $f(x)$, integration can then be regarded as an operation on sets, where the input is a set A and the output is its total mass:

$$A \to \int_A f(x)\mathrm{d}x \tag{39}$$

In this section we will generalize the above using measures. Just like in (39), a measure μ will take a set A as input and spit out a positive number — its total mass:

$$A \to \mu(A) \tag{40}$$

Before we rigorously define what constitutes a measure μ, we have to state the domain of μ, i.e. what the input of μ can be. As inputs to μ are subsets of Ω, the domain of μ will be a set of sets. In other words, we need to state the set of all sets A for which we can measure their mass $\mu(A)$. It seems feasible to simply assume that any set $A \subset \Omega$ can be assigned a measure. Or in technical terms, to assume that the domain of μ is the power set $\mathcal{P}(\Omega) = \{A : A \subset \Omega\}$. However, as we will see later such an assumption is very restrictive for the analysis of measures. We will therefore be interested in a bigger class of domains called σ-algebras, which are defined as follows.

Definition 8. *Let Ω be a set and $\mathcal{F} \subset \mathcal{P}(\Omega)$ a set of its subsets. We call $\mathcal{F}$ a σ-algebra if:*

- $\Omega \in \mathcal{F}$
- $A \in \mathcal{F} \implies A^c \in \mathcal{F}$
- *If $A_n \in \mathcal{F}$ for all $n \in \mathbb{N}$, then $\cup_{n\in\mathbb{N}} A_n \in \mathcal{F}$*

We will then call $(\Omega, \mathcal{F})$ a measurable space and any set $A \in \mathcal{F}$ a measurable set.

Let's decypher the above definition. The first point simply means that Ω has to be measurable, i.e. we can compute the mass of the whole set $\mu(\Omega)$. The second point means that if A is measurable then its complement A^c is also measurable. It makes sense, as we would expect that if Ω has mass $\mu(\Omega)$ and A has mass $\mu(A)$ then the mass of $A^c = \Omega/A$ should be available to us as $\mu(A^c) = \mu(\Omega) - \mu(A)$. The third point is the most delicate one. It states that if A_n is a *countable(!)* sequence of measurable sets, then their union is also measurable. That is, if I am able to measure some collection of sets, then I should be able to measure their union as well. While the collection is allowed to be infinite, it has to be countably infinite. For example, if $\mathcal{F}$ is a σ-algebra for $\Omega = \mathbb{R}$ and $\{a\} \in \mathcal{F}$ for all $a \in \mathbb{R}$, it *does not* mean that $\cup_{a\in[0,1]}\{a\} = [0,1] \in \mathcal{F}$. However, it does mean that $\{a, b\} \in \mathcal{F}$, or in fact, $\mathbb{Q} \in \mathcal{F}$, since the set of rational numbers is countably infinite.

Before you continue, make sure you are comfortable with the fact that $(\Omega, \mathcal{P}(\Omega))$ is a σ-algebra.

Now we are ready to state what a measure is.

Definition 9. *Let $(\Omega, \mathcal{F})$ be a measurable set and $\mu : \mathcal{F} \to [0, \infty]$ a function. We call μ a measure if:*

- $\mu(\emptyset) = 0$
- *If $A_n \in \mathcal{F}$ is a sequence of **disjoint** measurable sets then*

$$\mu(\cup_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} \mu(A_n) \tag{41}$$

We then say that $(\Omega, \mathcal{F}, \mu)$ is a measure space. If $\mu(\Omega) < \infty$ we say that μ is finite. Moreover, if $\mu(\Omega) = 1$ we call $(\Omega, \mathcal{F}, \mu)$ a probability space.

A measure space is a quite simple and intuitive concept — if A and B are both measurable and disjoint, then clearly $A \cup B$ should be measurable and its mass should be $\mu(A \cup B) = \mu(A) + \mu(B)$. The same logic

extends to a finite or countably infinite collection of measurable disjoint sets. Although the definitions of σ-algebras and measures are slightly cumbersome, there is logic to them.

In particular, suppose that we want to recover the classical notion of length, i.e. we set $\Omega = \mathbb{R}$ and seek to construct a measure μ such that $\mu([a,b]) = b - a$ for any interval $[a,b] \subset \mathbb{R}$. It turns out that if we insist that $\mathcal{F} = \mathcal{P}(\Omega)$, then there is no such measure. This means that if we want to construct a notion of measure that coincides with the classical notion of length, we have to concede the ability of measuring all subsets of $\mathbb{R}$. Therefore, the definition of a σ-algebra is crucial for the study of measures. [4]

So how does one define a σ-algebra on $\mathbb{R}$ for which a measure with $\mu([a,b]) = b-a$ would exist? The answer is long and technical, and we leave that aside. In this course we will work with the Borel σ-algebra, denoted as $\mathcal{B}(\mathbb{R})$. Technically speaking, $\mathcal{B}(\mathbb{R})$ is the smallest σ-algebra that contains all open sets. For us, this means that $\mathcal{B}(\mathbb{R})$ contains most of the sets that we want to measure. Using this σ-algebra, we can define the said measure $m_L : \mathcal{B}(\mathbb{R}) \to [0,\infty]$:

$$m_L(A) = \inf\left\{\sum_{n\in\mathbb{N}} b_n - a_n : a_n < b_n \text{ are such that } A \subset \cup_{n\in\mathbb{N}}[a_n, b_n)\right\} \tag{42}$$

This measure is called the Lebesgue measure. Checking that it is indeed a measure is a sweaty exercise, which we put aside.

Another classical example of a measure is the counting measure on the set of natural numbers. For a set A the counting measure simply states how many natural numbers are in A and can be defined on the measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ as

$$m_C(A) = \begin{cases} |A \cap \mathbb{N}|, & \text{if } A \cap \mathbb{N} \text{ is finite} \\ \infty, & \text{otherwise} \end{cases} \tag{43}$$

i.e. the number of elements in A. You should verify that $(\mathbb{R}, \mathcal{P}(\mathbb{R}), m_C)$ is a measure space.

We already mentioned that measure spaces with total mass 1 are called probability spaces, but how do we interpret them as such? A probability space $(\Omega, \mathcal{F}, \mathbb{P})$ represents an experiment. Ω is called the sample space and any element of $\omega \in \Omega$ is an outcome of the experiment. For example, suppose that our experiment involves flipping a fair coin twice. Then we can write $\Omega = \{HH, HT, TH, TT\}$. The σ-algebra $\mathcal{F}$ is called a family of events. Loosely speaking, each event describes whether some phenomenon happened or not in the experiment. For example, in the coin flip experiment the σ-algebra could be $\mathcal{F} = \mathcal{P}(\Omega)$ with the set $\{HH, HT, TH\} \in \mathcal{F}$ being the event "at least one coin shows heads". Finally, the measure $\mathbb{P}$ is the measure assigning probabilities to each event.

The intuition of such a setup is the following. Tyche, Goddess of Chance, chooses a point $\omega \in \Omega$ 'at random' according to the measure $\mathbb{P}$. Meaning, for $A \in \mathcal{F}$, $\mathbb{P}(A)$ represents the 'probability' (in the sense understood by our intuition) that the point ω chosen by Tyche belongs to A.[5]

Much more complex experiments can be described by measure spaces, however, more tools are required to move away from trivial examples. In addition, describing the underlying measure space $(\Omega, \mathcal{F}, \mathbb{P})$ can be rather cumbersome and is almost never done in practice. Instead measurable functions are used.

[4] A nice example of this is the Banach-Tarski paradox. It shows that a sphere of surface area 1 can be divided into 10 identical parts such that two spheres of sufrace area 1 can be constructed using 5 parts each. That is, we started off with a total surface area of 1 and ended up with a total surface area of 2. The key note here is that those 10 parts are non-measurable sets that do not preserve the total surface area as measurable sets would.

[5] Taken from David Williams book "Probability with Martingales".

2.3 Measurable functions

Suppose that in the above experiment we get a pound for each coin that turns heads. In other words, our reward is given by the function $R : \Omega \to \mathbb{R}$:

$$\begin{aligned} R(HH) &= 2 \\ R(HT) &= 1 \\ R(TH) &= 1 \\ R(TT) &= 0 \end{aligned}$$

Intuitively, a specific probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$, i.e. the probabilities of the coins turning heads or tails, and the reward function R should imply a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, i.e. the probabilities on the number of pounds earned.

More generally, let $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$ be two measurable spaces and consider a function $f : X \to Y$. If we have a measure μ_X on the space $(X, \mathcal{X})$ does f imply a measure μ_Y on the space $(Y, \mathcal{Y})$? If so, how do we construct such a measure? To assign a measure to a set $A \in \mathcal{Y}$ using the measure μ_X, we need to transform A to some set $B \in \mathcal{X}$. As f maps from X to Y and not the other way around, the only way to do so is to take the inverse $f^{-1}(A) = \{x \in X : f(x) \in A\}$. But for general functions f, $f^{-1}(A)$ is not guaranteed to be in $\mathcal{X}$. So the answer to the above depends on the function f.

Definition 10. *We say that a function $f : X \to Y$ between two measurable spaces $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$ is $(\mathcal{X}, \mathcal{Y})$-measurable (or simply measurable) if for all $A \in \mathcal{Y}$, $f^{-1}(A) \in \mathcal{X}$.*

If $X = Y = \mathbb{R}$ and $\mathcal{X} = \mathcal{Y} = \mathcal{B}(\mathbb{R})$ we will call such functions Borel measurable functions. If $(X, \mathcal{X})$ is a part of a probability space, f is also called a random variable.

Now is a good exercise for you to check that if f is measurable, then

$$\mu_Y(A) = \mu_X(f^{-1}(A)) \tag{44}$$

is a measure.

Definition 11. *Given a measure space $(X, \mathcal{X}, \mu)$, a measurable space $(Y, \mathcal{Y})$ and a measurable function $f : X \to Y$, the pushforward measure $f_{\#}\mu : \mathcal{Y} \to [0, \infty]$ is defined as*

$$f_{\#}\mu(A) = \mu(f^{-1}(A)) \tag{45}$$

When $(X, \mathcal{X}, \mu)$ is a probability space, the pushforward measure $f_{\#}\mu$ is called the law of the random variable f.

The significance of this is that a measure μ on $(X, \mathcal{X})$ and a measurable function $f : X \to Y$ together induce a measure on $(Y, \mathcal{Y})$. For example, in the above experiment the pushforward measure $R_{\#}\mathbb{P}$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is a measure describing the probabilities of possible rewards.

If μ is the Lebesgue measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $f(x) = 10x$, what kind of a measure is $f_{\#}\mu$?

It's important to mention that Borel measurable functions are closed under many operations. If f and g are Borel measurable functions, $\lambda \in \mathbb{R}$, then $f + g, fg, \lambda f$ are all Borel measurable functions. In addition, if f_n

is a sequence of Borel measurable functions, then $\inf f_n, \sup f_n, \liminf f_n, \limsup f_n$ are all Borel measurable functions too. Essentially, most functions that we know are Borel measurable, and coming up with non-Borel measurable functions is rather difficult.

Finally, before we go further, it is important to cover a bit of notation. Given measurable functions $f : X \to Y$ and a measure μ on $(X, \mathcal{X})$, we will often make statements about f. For example, we might say "$f(x) < 1$ on a set of measure 10". We would write this as:

$$\mu(\{x : f(x) < 1\}) = 10 \tag{46}$$

We will often use shorthands for such statements, e.g. $\mu(\{f < 1\}) = 10$, or even drop the braces. In probability spaces, we will often denote random variable by capitals and use the same notation. For example, in the above experiment, we write the probability of getting at least one pound as $\mathbb{P}(R \geq 1)$.

Moreover, we will see later that in measure theory whatever happens on a set of measure 0 simply does not matter. If a property of a measurable function holds everywhere, except on some set of measure 0, then we will say that it holds almost everywhere. For example, if the full measure of Ω is $\mu(\Omega) = 10$ and $\mu(\{f < 1\}) = 10$, then $\mu(\{f > 1\}) = 0$. We then say that $f < 1$ almost everywhere (a.e.). When $(\Omega, \mathcal{F}, \mu)$ is a probability space, we use the term almost surely (a.s.), or with probability 1.

Finally, in the case of probability spaces, describing $(\Omega, \mathcal{F}, \mathcal{P})$ can get rather cumbersome and is usually avoided. Instead, we say, for example, "R is distributed as a Bernoulli random variable". By this we mean that there is some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which X is a random variable and the law $\mu_X = X_{\#}\mathbb{P}$ of X is a Bernoulli distribution.

2.4 Lebesgue integral

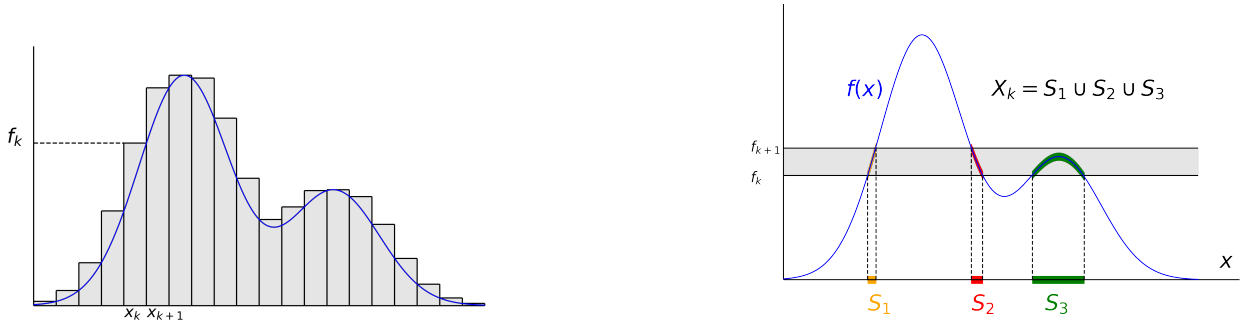
We are now well equipped to describe the Lebesgue integral. Figure 5 illustrates the conceptual difference between Riemann and Lebesgue integration. In Riemann integration, we approximate the integrand with step functions comprised of rectangles, for which the area is known. In Lebesgue integration, the conceptual difference is the observation that rectangles of the same height can be joined together into one shape. Then, the area covered by the shape is the height of those rectangles times the total *measure* of the intervals defining the rectangles.

For example, if $f : \mathbb{R} \to \mathbb{R}$ is a Borel measurable function, then the set $A = f^{-1}([y - \epsilon, y + \epsilon])$ is a Borel measurable set, which contains all points x such that $f(x) \in [y - \epsilon, y + \epsilon]$. On the set A we can approximate f by the constant value of y with an error of ϵ. Since A is a Borel measurable set, we can approximate the area of f on the set A by $\mu(A) \times y$, where μ is the Lebesgue measure. We can approximate f on the whole of $\mathbb{R}$ by partitioning the *range* of f into a set of intervals $\mathbb{R} = \cup_{n \in \mathbb{N}}[y_n - \epsilon, y_n + \epsilon]$, and approximating on each interval as above.

Notice that in this construction we moved from partitioning the domain, to partitioning the range, and from using the length of an interval, to using the measure of a set. This allows us to define integration on arbitrary measure spaces. We now look at how this is done rigorously.

Just like step functions were a basic building block in Riemann integration, simple functions will be a basic building block in Lebesgue integration.

Definition 12. *Let $(\Omega, \mathcal{F})$ be a measurable space. We say that $s : \Omega \to \mathbb{R}$ is a simple function, if for some measurable sets $E_1, \ldots, E_N \in \mathcal{F}$:*



(a) Riemann integral

(b) Lebesgue integral

Figure 5

$$s(x) = \sum_{i=1}^{N} c_i \mathbb{1}_{E_i}(x) \tag{47}$$

where $c_i \in \mathbb{R}$ are constants.

In addition, if μ is a measure on $(\Omega, \mathcal{F})$, we define the integral of a simple function w.r.t. μ as

$$\int_\Omega s(x)\mu(\mathrm{d}x) = \sum_{i=1}^{N} c_i \mu(E_i) \tag{48}$$

Notice that the definition is almost identical to a step function, we just swapped intervals $[t_{i-1}, t_i)$ to the more general measurable sets E_i, and exchanged $t_i - t_{i-1}$ to the measure of E_i.

We could now continue defining the integral just like in Riemann integration, but due to technical reasons the definition will be slightly different.

Definition 13. *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and $f : \Omega \to [0, \infty]$ a non-negative measurable function. We define the integral of f w.r.t μ as*

$$\int_\Omega f(x)\mu(\mathrm{d}x) = \inf_{s \geq f} \int_\Omega s(x)\mu(\mathrm{d}x) \tag{49}$$

where the supremum is over simple functions. If the integral is finite, we say that f is integrable.

If $f : \Omega \to \mathbb{R}$ is a measurable function, then its positive and negative parts, f_+ and f_-, defined as

$$f_+(x) = \max\{0, f(x)\} \tag{50}$$
$$f_-(x) = -\min\{0, f(x)\} \tag{51}$$
$$\tag{52}$$

are measurable functions such that $f(x) = f_+(x) - f_-(x)$. We say that f is integrable w.r.t. μ, if both f_+ and f_- are integrable. In that case, the integral of f is defined as

$$\int_\Omega f(x)\mu(\mathrm{d}x) = \int_\Omega f_+(x)\mu(\mathrm{d}x) - \int_\Omega f_-\mu(\mathrm{d}x) \tag{53}$$

When $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $X : \Omega \to \mathbb{R}$ is a random variable, we also write

$$\mathbb{E}[X] = \int_\Omega X(\omega)\mathbb{P}(\mathrm{d}\omega) \tag{54}$$

We will often use shorthands for Lebesgue integral notation. For example we will write $\int f\mathrm{d}\mu$ for $\int_\Omega f(x)\mu(\mathrm{d}x)$. And if the measure μ is the canonical Lebesgue measure, we will simply write $\int f\mathrm{d}x$ or $\int_\Omega f(x)\mathrm{d}x$. Sometimes Lebesgue integrals are also written as $\int_\Omega f(x)d\mu(x)$ (as in some slides in the lectures).

Lebesgue integral satisfies some basic properties. E.g. if f and g are both integrable and $\lambda \in \mathbb{R}$, then $\int f + \lambda g\mathrm{d}\mu = \int f\mathrm{d}\mu + \lambda \int g\mathrm{d}\mu$. Moreover, suppose further that $f(x) = g(x)$ almost everywhere, i.e. f and g disagree on a set of measure zero. Then, because sets of measure 0 do not matter, we have that $\int f\mathrm{d}\mu = \int g\mathrm{d}\mu$.

An interesting observation is that the Lebegsue integral of $f(x)$ with the counting measure m_C is simply a sum:

$$\int_\Omega f(x)m_C(\mathrm{d}x) = \sum_{n\in\mathbb{N}} f(n) \tag{55}$$

A couple of very important convergence theorems hold for the Lebesgue integral:

Theorem 3 (Monotone Convergence Theorem)**.** *Let $f_n(x)$ be a sequence of integrable functions such that $f_n(x) \le f_{n+1}(x)$ almost everywhere, and $\sup_n \int_\Omega f_n(x)\mu(\mathrm{d}x) < \infty$. Then the pointwise limit $f(x) = \lim_{n\to\infty} f_n(x)$ is also an integrable function with*

$$\int_\Omega f(x)\mu(\mathrm{d}x) = \lim_{n\to\infty}\int_\Omega f_n(x)\mu(\mathrm{d}x) \tag{56}$$

Theorem 4 (Dominated Convergence Theorem)**.** *Let $f_n(x)$ be a sequence of integrable functions. Suppose that there is an integrable function $g(x)$ such that $|f_n(x)| \le g(x)$ almost everywhere. Then the pointwise limit $f(x) = \lim_{n\to\infty} f_n(x)$ is also an integrable function with*

$$\int_\Omega f(x)\mu(\mathrm{d}x) = \lim_{n\to\infty}\int_\Omega f_n(x)\mu(\mathrm{d}x) \tag{57}$$

We saw that $f_n(x) = x^n$ is converges pointwise but not uniformly to the function $f(x) = 0$ on the interval $[0, 1)$, and so we could not exchange the limit with the Riemann integral. However, Monotone Convergence Theorem allows us to exchange the limit with the Lebesgue integral.

2.5 Radon-Nikodym derivatives

Now that we have defined a measure and Lebesgue integral it is a good time to look back at equation (39). We mentioned that the integral of the density $f(x)$ over the set A should represent its mass. Indeed, suppose that $A \in \mathcal{F}$. We define the integral of f over the set A as $\int_A f(x)\mu(\mathrm{d}x) = \int_\Omega \mathbb{1}_A(x)f(x)\mu(\mathrm{d}x)$. Now a natural question to ask is, if $f(x)$ is non-negative, is the map $\nu : \mathcal{F} \to [0, \infty]$ defined as

$$\nu(A) = \int_A f(x)\mu(\mathrm{d}x) \tag{58}$$

a measure? The answer is yes and it's again a good exercise to try and verify it with Monotone Convergence Theorem. In addition, if we swapped f for a function g, such that $f = g$ almost everywhere, the measure ν would not change.

The above shows that if we have a measure μ and a non-negative function f, we can define another measure ν, with f being the "connection" between the two measures. Conversely, given two measures μ and ν, will there always be such a function f?

Firstly, note that if $\mu(A) = 0$, then we have $\int_A f(x)\mu(\mathrm{d}x) = 0$. That is, $\mu(A) = 0$ has to imply $\nu(A) = 0$. If it does, we have the following result.

Definition 14. *Let μ and ν be two measures on a measurable space $(\Omega, \mathcal{F})$. We say that ν is absolutely continuous w.r.t. μ, written $\nu \ll \mu$ if for any $A \in \mathcal{F}$*

$$\mu(A) = 0 \implies \nu(A) = 0 \tag{59}$$

Moreover, we say that μ and ν are equivalent if $\mu \ll \nu$ and $\nu \ll \mu$.

Theorem 5 (Radon-Nikodym Theorem)**.** *Let μ and ν be two finite[6] measures on a common measurable space $(\Omega, \mathcal{F})$. If $\nu \ll \mu$, then there exists a measurable function $f(x)$ such that for all $A \in \mathcal{F}$ equation (58) holds. f is unique almost everywhere and is called the Radon-Nikodym derivative of ν w.r.t. μ. The following notation is used:*

$$f = \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \tag{60}$$

$$\nu(dx) = f(x)\mu(\mathrm{d}x) \tag{61}$$

Theorem 6. *Let μ and ν be measures on $(\Omega, \mathcal{F})$, such that $\nu \ll \mu$, and let $f = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}$. Then for a measurable function g, we have that*

$$\int_\Omega g(x)\nu(\mathrm{d}x) = \int_\Omega g(x)f(x)\mu(\mathrm{d}x) \tag{62}$$

For example, consider a measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ that puts a uniform measure on the interval $[0, 1]$ only:

$$\mu_1(A) = m_L(A \cap [0, 1]) \tag{63}$$

Then you should check that $\mu_1 \ll m_L$ and that the Radon-Nikodym derivative is given by $f_1(x) = \mathbb{1}_{[0,1]}(x)$.

Or, if the measure is defined as

$$\mu_2(A) = \begin{cases} 2, & \text{if } 1, 2 \in A \\ 1, & \text{if } 1 \in A \text{ or } 2 \in A \\ 0, & \text{otherwise} \end{cases} \tag{64}$$

[6] the theorem holds for more general σ-finite measures as well

then $\mu_2 \ll m_C$ and the Radon-Nikodym derivative is given by

$$f_2(x) = \begin{cases} 1, & x = 1, 2 \\ 0, & \text{otherwise} \end{cases} \tag{65}$$

Conversely, note that both measures could have been defined through the Radon-Nikodym derivative $f(x)$ as

$$\mu(A) = \int_A f(x) m_L(\mathrm{d}x) \tag{66}$$

or m_C instead of m_L.

In addition, note that if we wanted to integrate a function $g(x)$ w.r.t. the measure μ_2 we could apply the above theorem:

$$\int_{\mathbb{R}} g(x)\mu_2(\mathrm{d}x) = \int_{\mathbb{R}} f_2(x)g(x)m_C(\mathrm{d}x) = \sum_{n\in\mathbb{N}} f_2(n)g(n) = f_2(1)g(2) + f_2(2)g(2) \tag{67}$$

Definition 15. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ a random variable with the law*

$$\mu_X(A) = \mathbb{P}(X^{-1}(A)) \tag{68}$$

If μ_X is absolutely continuous w.r.t. the Lebesgue measure m_L (counting measure m_C) we call X a continuous random variable and define it's probability density (mass) function as the Radon-Nikodym derivative $\frac{\mu_X}{m_L}$ ($\frac{\mu_X}{m_C}$).

That is, when we say "X is distributed as a standard Gaussian random variable", we mean that there is some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which X is a random variable, and the law μ_X of X is Gaussian. That is to say, μ_X is absolutely continuous w.r.t. Lebesgue measure m_L and has a Radon-Nikodym derivative given by:

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) \tag{69}$$

2.6 Product measure spaces

Often, if we have two independent experiments described by probability spaces $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$, e.g. flipping a coin and drawing a random number between $[0, 1]$, we might want to analyze the combination of the two experiments as described in the very first example of the chapter. This is done via product measure spaces.

Construction of product spaces is very technical, so we provide just a few details. The product space $(\Omega, \mathcal{F}, \mu)$ of $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$ is defined as follows. The sample space Ω is just all pairs between Ω_1 and Ω_2 — $\Omega = \Omega_1 \times \Omega_2$. The σ-algebra $\mathcal{F} = \mathcal{F}_1 \otimes \mathcal{F}_2$ is the smallest σ-algebra such that for all $A \in \mathcal{F}_1, B \in \mathcal{F}_2$, $A \times B \in \mathcal{F}$. And finally, the measure $\mu = \mu_1 \otimes \mu_2$ is a measure such that for all $A \in \mathcal{F}_1, B \in \mathcal{F}_2$

$$\mu(A \times B) = \mu_1(A)\mu_2(B) \tag{70}$$

For the purposes of integration, we have the following famous theorem.

Theorem 7 (Fubini-Tonelli theorem)**.** *Let $(X, \mathcal{X}, \mu_X)$ and $(Y, \mathcal{Y}, \mu_Y)$ be measure spaces with measures μ_X and μ_Y being σ-finite measures, and f a measurable function on the product space $(X \times Y, \mathcal{X} \otimes \mathcal{Y}, \mu_X \otimes \mu_Y)$. Then*

$$\int_X \int_Y |f(x,y)| \mu_Y(\mathrm{d}y) \mu_X(\mathrm{d}x) = \int_Y \int_X |f(x,y)| \mu_X(\mathrm{d}x) \mu_Y(\mathrm{d}y) = \int_{X \times Y} |f(x,y)| (\mu_X \otimes \mu_Y)(\mathrm{d}(x,y)) \quad (71)$$

In addition, if any of the above integrals are finite, f is integrable and we also have

$$\int_X \int_Y f(x,y) \mu_Y(\mathrm{d}y) \mu_X(\mathrm{d}x) = \int_Y \int_X f(x,y) \mu_X(\mathrm{d}x) \mu_Y(\mathrm{d}y) = \int_{X \times Y} f(x,y) (\mu_X \otimes \mu_Y)(\mathrm{d}(x,y)) \quad (72)$$