Lecture 1: Review of probability theory / Introduction to Stochastic processes

Readings You should make sure you are comfortable with the following concepts from probability theory:

- probability space
- random variable
- expectation, integration with respect to a probability measure
- conditional probability and conditional expectation
- Law of Total Probability
- independence
- convergence of random variables (optional)
- Gaussian random variables (optional, but recommended)

These are reviewed briefly in sections 1.3–1.5 of these notes. Here are some additional references to help review these concepts. These are suggested readings only; you should read whatever you find helpful to fill in your background, and you are welcome to find other sources.

- Pavliotis, Appendix B, p.303-319. Skip B.5.1, and B.4 is optional.
- Lecture notes 1, 2 from 2014 (posted to NYU classes)
- Appendix A.4 from Durrett *Probability: Theory and Examples*, for more details on integration with respect to a measure (posted to NYU classes)
- Koralov and Sinai (2010) Ch. 3 p. 37 for more details on integration with respect to a measure
- Koralov and Sinai (2010), Grimmett and Stirzaker (2001), Breiman (1992), and other probability books, for more details on probability spaces / random variables

Full references are at the end of these notes.

1.1 Course plan

The course will be divided roughly equally into two parts:

Part I will focus on *Stochastic processes*
Part II will focus on *Stochastic calculus*.

Today we will give an overview of the topics we will cover, and briefly review some probability theory.

1.2 Introduction: what is a stochastic process?

A *stochastic process* is a random function of a single variable, usually time. Specifically, let $(\Omega, \mathscr{F}, P)$ be a probability space (see Section 1.3), and let T be an ordered set, called the *index set*.

Definition. A *stochastic process* is a collection of random variables $\{X_t,\ t \in T\}$, such that for each $t \in T$, X_t is a random variable on $(\Omega, \mathscr{F}, P)$.

We will often write a stochastic process as $(X_t)_{t\in T}$ or simply X_t, when it is clear from the context what the index set is and that we refer to the process, not a particular random variable.

Definition. When the index set is countable the process is called *discrete-time*. Examples include $T = \mathbb{Z}^+$, $T = \{0, 1, \ldots, N\}$, etc. When the index set contains an interval of $\mathbb{R}$ the process is called *continuous-time*. Examples include $T = \mathbb{R}$, $T = [a, b]$, etc. An index set may also be a mixture of continuous and discrete parts.

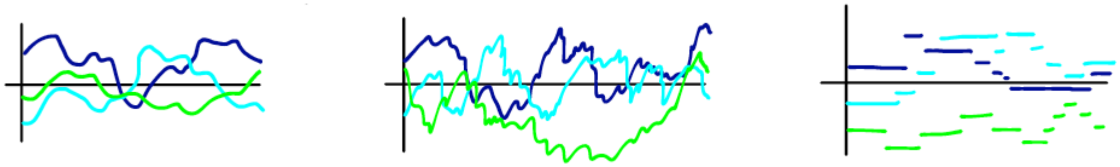
To highlight its two arguments, a stochastic process may be written as $X(t, \omega)$ or $X_t(\omega)$ or $X(t)(\omega)$, where $t \in T$ and $\omega \in \Omega$.

We can think of a stochastic process in two ways:

(1) A parameterized family of random variables, given by the function $t \to X_t(\omega)$.
At each fixed t, we have a collection of random variables.

(2) A probability distribution on path space, given by the function $\omega \to X_t(\omega)$.
Each realization of the randomness, $\omega \in \Omega$, gives a function of $t \in T$. Events are sets of paths, e.g. "all paths that pass through $[x, x+h]$ at time t_0", "all paths whose maximum value is bigger than M", etc. We cannot construct a probability "density" on this space, because the space is infinite-dimensional. However, we can still talk about probabilities of events.

Here are some examples of realizations of stochastic processes:



Stochastic processes are used as models for a wide range of phenomena. Examples include:

- velocity of a particle in a smooth chaotic flow
- position of grain of pollen in a glass of water
- mean temperature of the planet
- price of a stock
- number of bees in a bee hive
- # of molecules in the reactant state of a chemical reaction
- bond angles of a folding protein
- randomly perturbed harmonic oscillator (swingset)
- size of a self-assembling viral capsid
- uncertainty in a weather prediction, depending on days in the future
- etc. You can think of many more.

Many of these examples are not inherently random. For example, we could in principle write down the Navier-Stokes equations describing the flow of a turbulent fluid, and solve them to very high precision to obtain the velocity at each location in the fluid. But, even for deterministic phenomena, it still makes sense to model them stochastically in many situations. There are several reasons for this:

- They could be very hard or impossible, to simulate to high enough precision.
- Even if we can simulate them, that doesn't necessarily give us useful information. For example, suppose we know where every atom is in a river, at every point in time. What have we learnt about the river? It is often more helpful to understand patterns, or large-scale, average, emergent behaviour, which is often best described statistically.
- The equations may be chaotic, i.e. very sensitive to small perturbations. If we don't know the exact initial condition then the uncertainty in the solution can be modelled stochastically.
- Often stochastic models are much more tractable, analytically and numerically, than their deterministic counterparts, so it can be useful to approximate a deterministic system with a stochastic one.
- Some systems are inherently stochastic, like quantum mechanics, or uncertainty (if you are a Bayesian.)

Note that "stochastic" $\neq$ "unstructured". A stochastic process can have a lot of structure. Our goal in this course is to understand and quantify this structure. We will be interested both in what we can say about the process analytically, and also how we can simulate it.

We will study several kinds of stochastic processes:

- Markov chains.
  We will spend some time studying Markov chains, both discrete-time and continuous-time. These are widely used as models, and additionally they are important for this course because many of the ideas used to understand Markov chains can be directly generalized to study diffusion processes (see below.) Also, one of the century's most powerful algorithms, Markov Chain Monte Carlo, is based on Markov chains, and we will study this in the third lecture. We will not study Markov chains exhaustively (you could take a whole course on them), but rather focus on ideas that we will see later in the course when we study stochastic differential equations.

- Gaussian processes
  These are models for a wide range of phenomena, even smooth ones, such as velocities or waves in a fluid, the height of topography, etc. They are studied because they have many statistical properties which are analytically tractable, and because they are expected to occur generically whenever a process is the result of a superposition of random effects, because of the Central Limit Theorem.

- Stationary processes
  Stationary processes are (roughly) those whose statistics do not change with time. They are studied when one is interested in understanding statistical "equilibrium," or steady-state. Stationary Gaussian processes are studied particularly widely and are easy to simulate. We will look at the stochastic analogue of the Fourier transform, and you will learn how to calculate the Fourier transform of functions which do not even live in L_1!

- Brownian motion
  This is perhaps the most famous stochastic process. It was originally invented to model the motion of pollen grains, but now the basis of Stochastic Calculus.

- Diffusion processes
  These are processes that are solutions to a stochastic differential equation, a stochastic analogue of an ordinary differential equation. They are our focus in Part II of the course. Specifically, we we study:

  - How can we make sense of an equation with randomness in it? For example, how could we make sense of this equation:

$$\underbrace{\frac{dx}{dt}}_{\substack{\text{position of}\\ \text{particle in}\\ \text{a fluid}}} = \underbrace{u(x,t)}_{\substack{\text{velocity of}\\ \text{fluid}}} + \underbrace{\eta(t)}_{\substack{\text{noise term,}\\ \text{e.g. unresolved velocity}}} \quad .$$

    or this one:

$$\underbrace{\frac{dN}{dt}}_{\substack{\text{rate of change}\\ \text{of population}}} = \Big(\underbrace{a(t)}_{\substack{\text{average}\\ \text{growth}\\ \text{rate}}} + \underbrace{\eta(t)}_{\text{noise}}\Big)\underbrace{N(t)}_{\substack{\text{population}\\ \text{size}}} \quad .$$

    Usually we want the noise term to be "as random as possible," say completely independent from timestep to timestep. It turns out this means we must choose the noise to be the "derivative" of Brownian motion, and furthermore, that the Riemann integral that we usually use for defining solutions to ODEs, is no longer well-defined. So, we will start with the fundamental issue of defining a new kind of integral, and then show this new integral gives rise to new rules of calculus, for example, the chain rule will change.

  - How can we solve these equations? Both analytically, and also numerically. For the latter, how can we measure how "close" we are to the true solution? We will see that adding randomness, also adds many more ways of measuring "closeness".

  - How can we solve for statistical properties of the solution, such as its probability density at some time, or the average time it takes to reach a boundary? We can answer questions like these by solving elliptic PDEs. We will spend a fair amount of time studying the link between diffusion processes and PDEs, since from the PDEs we can use an arsenal of applied math techniques to make approximations when we can't solve equations exactly.

- Detailed balance
  Throughout the course, we will pay attention to whether the processes we study are physically reasonable, i.e. whether they satisfy a condition know as detailed balance, when appropriate. Roughly, we know that the laws of physics look "the same" forwards and backwards in time, so our stochastic models should too.

This list is certainly not exhaustive. In particular, we will not study jump processes, point processes, or Levy processes, although these processes arise in many modelling situations. However, after this course you should have the tools to learn about these processes on your own.

The approach in this course will be to study these objects from an applied math perspective. This means that we will focus on learning tools to solve concrete problems, on issues that are important in physically-motivated problems, and on simulation techniques. It also means that we won't necessarily prove every statement we make. We will be rigorous when it makes sense to be, and doesn't detract too much from studying typical versions of these objects. However, we won't always prove the mathematical objects we study actually exist, we won't worry about pathological cases that don't arise in applications, and we will assume all functions are as smooth as is necessary to perform the calculations we need. We will avoid measure theory, and functional analysis, as much as possible, but be warned that to make these notes rigorous, you could spend an entire course studying the subtle details that we will gloss over.

1.3 A brief review of some probability theory

Definition. A *probability space* is a triple $(\Omega, \mathscr{F}, P)$, where

1. Ω is the *sample space*, i.e. the set of all possible outcomes. Each element $\omega \in \Omega$ is a *sample point*. A subset $A \subset \Omega$ is an *event*.
2. $\mathscr{F}$ is a σ-field of subsets of Ω, i.e. it satisfies[1]
   (a) $\Omega \in \mathscr{F}, \emptyset \in \mathscr{F}$;
   (b) $A \in \mathscr{F} \Rightarrow A^c \in \mathscr{F}$;
   (c) $A_1, A_2, \ldots \in \mathscr{F} \Rightarrow \cup_{n=1}^{\infty} A_n \in \mathscr{F}$.
3. P is a *probability measure*, i.e. a set function which satisfies
   (a) $P(\emptyset) = 0$, $P(\Omega) = 1$;
   (b) If $A_1, A_2, \ldots$ are pairwise disjoint ($A_i \cap A_j = \emptyset$ if $i \neq j$), then $P(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$.

We won't worry too much about $\mathscr{F}$ in this course; we'll assume it allows us to speak about what we want, which is almost always true in applications.

Examples

1. Toss a fair coin n times. Then

$$\begin{aligned}\Omega &= \{\text{HH}\ldots\text{H, HH}\ldots\text{,T,}\ldots\text{,TT}\ldots\text{T}\} = \{\text{H,T}\}^n = \text{all strings of H,T of length } n.\\ \mathscr{F} &= \text{all subsets of } \Omega\\ P(A) &= \frac{|A|}{2^n}.\end{aligned}$$

For example, for $n = 4$, we would have $P(\{\text{HHHT,HHTH,HHTT}\}) = 3/16$.

2. Spin a top, and measure its angle with North when it falls. Then

$$\begin{aligned}\Omega &= [0, 2\pi)\\ \mathscr{F} &= \text{Borel sets} \cap [0, 2\pi)\\ P(A) &= \frac{\mu(A)}{2\pi}, \quad \text{where } \mu = \text{Lebesgue measure, i.e. } \mu([a,b]) = b - a..\end{aligned}$$

Definition. A *random variable* is a function $X : \Omega \to S \subset \mathbb{R}$ such that X is $\mathscr{F}$-measurable,[2] i.e. $X^{-1}(B) \in \mathscr{F}$ for all Borel sets $B \subset \mathbb{R}$. The set of possible values S is called the *state space*.

Remark. A random variable X induces a probability measure μ on $\mathbb{R}$, given by

$$\mu(B) = P(X \in B) = P(X^{-1}(B)).$$

The proof is straightforward and is an ELFS.[3]

[1] If you don't know what a σ-field is, then don't worry about this; it is included merely for completeness. In applications, you can think of $\mathscr{F}$ as the set of all events that you might possibly dream of asking about.

[2] Again, if you haven't seen this before it doesn't matter.

[3] Exercise Left For Student.

Because of this remark, we often just work with random variables and the measure they induce, and forget about the probability space underlying them.

Definition. The *distribution function* or *cumulative distribution function* of a random variable X is the function $F:\mathbb{R}\to[0,1]$ given by $F(x)=P(X\leq x)$.

Remark. Recall that a distribution function satisfies (i) $\lim_{x\to-\infty}F(x)=0$, $\lim_{x\to\infty}F(x)=1$, (ii) $F(x)$ is monotonically increasing, and (iii) $F(x)$ is right-continuous, i.e. $\lim_{h\to 0}F(x+h)=F(x)$.

Definition. X is *discrete* if it takes values in some countable subset $S=\{x_1,x_2,\ldots\}$ of $\mathbb{R}$. A discrete random variable has a *probability mass function* (pmf) $f:\mathbb{R}\to[0,1]$ given by $f(x)=P(X=x)$. We can calculate probabilities of events $A\subset S$ as $P(X\in A)=\sum_{x_i\in A}f(x_i)$.

Definition. X is *continuous* if there is an integrable function $f:\mathbb{R}\to[0,\infty)$ such that $F(x)=\int_{-\infty}^{x}f(u)du$. If this is the case, then f is called the *probability density function* (pdf) of X. We can calculate probabilities of events as $P(X\in A)=\int_A f(u)du$.

Remark. Not all random variables are continuous or discrete. They can be a mixture, or they can have a singular part, as for the Cantor measure. The full classification is given by Lebesgue's Decomposition Theorem.

Definition. The *expectation* of a random variable X is defined as

$$\mathbb{E}X=\int_\Omega X(\omega)\,P(d\omega)=\int_{\mathbb{R}}x\,dF(x)=\begin{cases}\int_{\mathbb{R}}xf(x)dx & X\text{ is continuous}\\ \sum_{x\in S}xf(x) & X\text{ is discrete}\end{cases}\tag{1}$$

This definition extends naturally to functions of X as $\mathbb{E}h(X)=\int h(x)dF(x)$.

Note that the definition above uses the notation $\int h(x)dF(x)$ to represent both a continuous integral and a discrete sum. This notation is convenient because then we don't need to worry about whether the random variable under consideration is continuous or discrete. If you are not familiar with this kind of integral, you should review Section 1.5 and the references within.

Definition. The *variance* of X is

$$\mathrm{Var}(X)\equiv\mathbb{E}(X-\mathbb{E}X)^2=\mathbb{E}X^2-(\mathbb{E}X)^2.$$

These definitions extend naturally to collections of random variables, as we summarize briefly below.

Definition. $X=(X_1,\ldots,X_n)^t\in\mathbb{R}^n$ in a *random vector* or *vector-valued random variable* if each component is a random variable. Then we define

(i) The mean, $\mathbb{E}X=(\mathbb{E}X_1,\ldots,\mathbb{E}X_n)^t$.
(ii) The covariance matrix, $\mathrm{cov}(X,X)=\mathbb{E}(X-\mathbb{E}X)(X-\mathbb{E}X)^t$.

Remark. The covariance is a matrix, with components $\mathrm{cov}(X,X)_{ij}=\mathrm{cov}(X_i,X_j)=\mathbb{E}(X_i-\mathbb{E}X_i)(X_j-\mathbb{E}X_j)^t$. This is just the covariance between the random variables X_i,X_j. For example, if $X=(X_1,X_2)^t$, then

$$B=\mathrm{cov}(X,X)=\begin{pmatrix}\mathrm{var}(X_1) & \mathrm{cov}(X_1,X_2)\\ \mathrm{cov}(X_1,X_2) & \mathrm{var}(X_2)\end{pmatrix}.$$

Clearly the covariance matrix B is positive semi-definite (ELFS), a property we will exploit heavily in lectures to come.

Definition. The *joint distribution function* is $F(x_1,\dots,x_n) = P(X_1 \le x_1,\dots,X_n \le x_n)$. A continuous vector-valued random variable has a *probability density function (pdf)* $f(x_1,\dots,x_n)$ which is such that $P(X \in A) = \int_A f(x_1,\dots,x_n)dx$, and a discrete vector-valued random variable has a *probability mass function (pmf)*, defined in a similar way. Sometimes these are called joint pdfs or joint pmfs, to emphasize that they describe the distribution of several random variables jointly.

Definition. Given a continuous random vector $X = (X_1,\dots,X_n)$, the *marginal pdf* of a component X_i is

$$f_{X_i}(x_i) = \int f(x_1,\dots,x_n)dx_1\cdots dx_{i-1}dx_{i+1}\cdots dx_n\,.$$

That is, we integrate the joint pdf over all variables except the variable of interest. A similar definition holds for the *marginal pmf* of a discrete random vector, with the integral replaced by a sum.

Important questions to ask about random variables are about independence, and also about conditional probabilities.

Definition. Two events $A,B \in \mathscr{F}$ are *independent* if $P(A\cap B) = P(A)P(B)$. Two random variables X,Y are *independent* if, for all Borel sets A,B, the events $X^{-1}(A)$ and $Y^{-1}(B)$ are independent, i.e. $P(X^{-1}(A)\cap Y^{-1}(B)) = P(X^{-1}(A))P(Y^{-1}(B))$. A collection of random variables $X_1,\dots,X_n$ is independent if for any Borel sets B_j, $P\big(\cap_{j=1}^n X_j^{-1}(B_j)\big) = \prod_{j=1}^n P(X_j^{-1}(B_j))$.

Remark. Continuous or discrete random variables X,Y are independent iff the joint pdf or pmf factors, as $f(x,y) = f_X(x)f_Y(y)$. A similar statement is true for a larger collection of random variables.

Definition. Given events A,B, the *conditional probability of A given B* is

$$P(A|B) \equiv \frac{P(A\cap B)}{P(B)}. \tag{2}$$

Remark. Events A,B are independent iff $P(A|B) = P(A)$.

A very useful theorem is the following:

Law of Total Probability (LOTP). Let $B_1,\dots,B_k$ be a sequence of events that *partition* the sample space: B_i are disjoint and their union is Ω. Then, for any event A,

$$P(A) = \sum_{i=1}^k P(A\cap B_i) = \sum_{i=1}^k P(A|B_i)P(B_i)\,. \tag{3}$$

Example (Dobrow (2016) p.11) According to the Howard Hughes Medical Institute, about 7% of men and 0.4% of women are colorblind – either cannot distinguish red from green or see red and green differently from most people. In the United States, about 49% of the population is male and 51% female (this data simplifies by ignoring other gender categories). A person is selected at random. What is the probability they are colorblind?

Solution Let C, M, F denote the events that a random person is colorblind, male, and female respectively. By the LOTP,

$$\begin{aligned} P(C) &= P(C|M)P(M) + P(C|F)P(F) \\ &= (0.07)(0.49) + (0.004)(0.51) = 0.03634. \end{aligned}$$

We will use various other versions of conditional probability. Often we will need to condition on the value of a random variable, for example as $P(Y \in A|X = x)$ where $A \subset S$ is some event in the state space. When X is discrete this is not a problem as we can simply define the conditional probability mass function to be

$$f(y|x) \equiv \frac{P(Y = y \cap X = x)}{P(X = x)} = \frac{f(x,y)}{f_X(x)} , \tag{4}$$

for all x such that $P(X = x) \neq 0$, and define it to be 0 otherwise. Here $f(x,y)$ is the joint pmf, and $f_X(x) = \sum_y f(x,y)$ is the marginal pmf of X. We can then calculate the desired conditional probability using $f(y|x)$ as the pmf.

For a continuous random variable X this is a problem because $P(X = x) = 0$ for all x. However, we can define the conditional density to be

$$f(y|x) = \frac{f(x,y)}{f_X(x)} , \tag{5}$$

where $f_X(x) = \int f(x,y)dy$ is the marginal density of X. This definition is equivalent to (4) in a limiting sense, if we take an appropriate decreasing sequence of set that shrink to the sets $\{X = x\}$, $\{Y = y\}$. For example, let us define $B_n = \{x \leq X \leq x+h_n\}$, $C_n = \{y \leq Y \leq y+h_n\}$, where $h_n \to 0$ as $n \to \infty$. Then the conditional probability $P(Y \in C_n|X \in B_n)$ approaches (5) as $n \to \infty$. We can then use $f(y|x)$ as a pdf when calculating probabilities of the form $P(Y \in A|X = x)$.

One must be careful when performing this limiting procedure as not all sequences of shrinking sets will give the same answer. To be fully rigorous we should condition on a σ-algebra, not a random variable. That is, we need to define the conditional expectation $\mathbb{E}(X|\mathscr{G})$ where $\mathscr{G}$ is a σ-algebra. Then, we can define the random variable $\mathbb{E}(Y|X)$ to be the random variable one obtains by conditioning on the σ-algebra generated by X. This is beyond the scope of this course so we will not worry about these issues, but please talk to the instructor if you would like to learn more about the possible pathologies and how to fix them.

1.4 Some useful distributions

You should be familiar with all of these distributions.

Uniform on $[a,b]$ (continuous)

$$f(x) = \frac{1}{b-a} 1_{[a,b]}(x), \qquad F(x) = \begin{cases} 0 & x \leq a \\ \frac{b-a}{x} & a \leq b \leq x \\ 1 & x > b \end{cases}$$

You can also define a uniform distribution on a discrete state space, in a similar way.

Bernoulli distribution, parameter p (discrete)

A Bernoulli random variable is like the result of a coin toss which comes up heads with probability p (a "success"), and tails with probability $1-p$ ("fail".)

$$f(0) = 1-p, \quad f(1) = p, \qquad \mathbb{E}X = p, \quad \text{Var}(X) = p(1-p).$$

Binomial distribution, parameters n, p (discrete)

This gives the probability of getting a certain number of heads in n coin tosses with parameter p, i.e. the probability of each number of successes in a sequence of n Bernoulli trials.

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0,1,\ldots,n. \qquad \mathbb{E}X = np, \quad \mathrm{Var}(X) = np(1-p).$$

Poisson distribution, parameter λ (discrete)

This models the number of events that occur in an interval of time, when the events occur completely independently of each other. E.g., the number of radioactive atoms that decay in an hour, the number of meteorites that hit the earth in a millenium, the number of deaths by horse kicks per year in the Prussian army, etc.

$$f(k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0,1,2,\ldots. \qquad \mathbb{E}X = \lambda, \quad \mathrm{Var}(X) = \lambda.$$

Exponential distribution, parameter λ (continuous)

Often used to model the waiting time between events that occur independently of each other, e.g. the time between which two atoms decay radioactively.

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0. \qquad \mathbb{E}X = \lambda, \quad \mathrm{Var}(X) = \lambda.$$

The exponential distribution is the only continuous distribution which describes a random variable X that is *memoryless*: $P(X > s+t | X > t) = P(X > s)$ for all $s,t > 0$.

Normal distribution, parameters μ, σ (continuous)

Also called the Gaussian distribution, or Bell curve. This is canonical because it occurs whenever there is a sum of a large number of (roughly) independent random variables, by the Central Limit Theorem.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \quad x \in \mathbb{R}. \qquad \mathbb{E}X = \mu, \quad \mathrm{Var}(X) = \sigma^2.$$

1.5 Integration with respect to to a measure / Expectation

We will occasionally need to integrate a function with respect to a measure. Here is a brief review of how to define such an integral. This follows Appendix A.4 from Durrett (2005); see this reference for more details, including proofs that the integral is well-defined.

Let $(\Omega, \mathscr{F}, \mu)$ be a measure space with (σ-finite) measure μ, and let $f : \Omega \to \mathbb{R}$ be a measurable function defined on the sample space. We wish to define the integral

$$\int f d\mu, \tag{6}$$

for a suitable class of measurable functions f. (Sometimes we write this as $\int f(x)\mu(dx)$.) Sometimes Ω is a sample space and μ is a probability measure, in which case $\int f d\mu = \mathbb{E}f$ (recall (1)).

Let's first consider some examples of measures μ. For all of the below, we will assume that $\Omega = \mathbb{R}$ and $\mathscr{F}$ is the Borel sets.

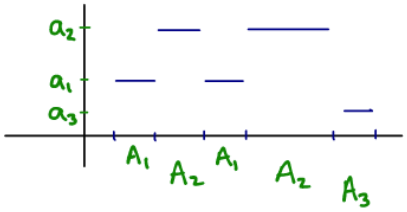
- μ is the Lebesgue measure, i.e. $d\mu(x) = dx$. Then, the construction below is exactly the Lebesgue integral, and we obtain
$$\int f d\mu = \int f(x)dx$$
where the second integral is the Lebesgue integral.
- μ is absolutely continuous with respect to the Lebesgue measure. Then we can write either $d\mu(x) = g(x)dx$ for some integrable function g (specifically, $g \in L^1(\mathbb{R})$), or $d\mu(x) = dG(x)$ where $G(x) = \int^x g dx$ is the antiderivative of g. We will often use this notation when G is the cumulative distribution function of a probability. The integral becomes
$$\int f d\mu = \int f dG = \int f(x)g(x)dx.$$
- μ is supported on a countable set of points $\{x_1, x_2, \dots\}$, so we can formally write $\mu(dx) = b_1\delta(x_1)dx + b_2\delta(x_2)dx + \cdots$, for $b_i \in \mathbb{R}^+$. The integral is
$$\int f d\mu = b_1 f(x_1) + b_2 f(x_2) + \cdots.$$

In general we will often use this construction when we have a cumulative distribution function of a probability G (recall (1)), so that $\mu([a,b]) = G(b) - G(a)$. We write $d\mu(x) = dG(x)$. If G is differentiable, then $d\mu(x) = G'(x)dx$. The integral is
$$\int f d\mu = \int f dG = \mathbb{E}f.$$

A rough definition of the integral in (6) comes from approximating it by the discrete sum over a partition $x_0, x_1, \dots, x_N$ as
$$\int_{\mathbb{R}} x\, dF(x) \quad \approx \quad \sum_i x_i\big(F(x_{i+1}) - F(x_i)\big)\Delta x_i, \qquad \Delta x_i = x_{i+1} - x_i.$$

To properly define the integral in (6) there are four steps, outlined below.

1. Let a *simple function* be a function of the form $\phi(x) = \sum_{i=1}^n a_i 1_{A_i}(x)$, where $a_i \in \mathbb{R}$, and $A_i \subset \Omega$ are disjoint sets that are measurable with respect to $\mathscr{F}$, and such that $\mu(A_i) < \infty$. The function $1_A(x)$ is an indicator function, i.e. $1_A(x) = 1$ if $x \in A$, $1_A(x) = 0$ if $x \notin A$. Here is an example:



Define the integral for a simple function to be

$$\int \phi(x) d\mu(x) = \sum_{i=1}^{n} a_i \mu(A_i) \,.$$

One can verify the integral above satisfies the following properties (see Durrett (2005) for the proofs):

(i) If $\phi \geq 0$ a.e.[4] then $\int \phi d\mu \geq 0$.
(ii) For any $a \in \mathbb{R}$, $\int a\phi d\mu = a \in \phi d\mu$.
(iii) $\int (\phi + \psi) d\mu = \int \phi d\mu + \int \psi d\mu$.
(iv) If $\phi \leq \psi$ a.e., then $\int \phi d\mu \leq \int \psi d\mu$.
(v) If $\phi = \psi$ a.e., then $\int \phi d\mu = \int \psi d\mu$.
(vi) $|\int \phi d\mu| \leq \int |\phi| d\mu$.

2. Let f be a bounded function that is non-zero only on a set $E \subset \Omega$ with $\mu(E) < \infty$. Define

$$\int f d\mu = \sup_{\substack{\phi \leq f \\ \phi \text{simple}}} \int \phi d\mu = \inf_{\substack{\psi \geq f \\ \psi \text{simple}}} \int \psi d\mu \,.$$

To show this is well-defined, one must show the sup and inf are equal. To do this one uses the properties (i)–(vi) above. Then, one can show that properties (i)–(vi) still hold, for bounded functions.

3. Let $f \geq 0$. Define

$$\int f d\mu = \sup\{\int h d\mu : \; 0 \leq h \leq f, \; h \text{ is bounded}\} \,.$$

This is well-defined already; we don't need to check anything. One can then show properties (i)–(vi) hold for nonnegative functions.

4. Suppose $\int |f| d\mu < \infty$. Let

$$f^+(x) = f(x) \vee 0 \quad \text{and} \quad f^-(x) = (-f(x)) \vee 0 \,,$$

where $a \vee b = \max(a,b)$. Then $f(x) = f^+(x) - f^-(x)$, and $|f(x)| = f^+(x) + f^-(x)$. Define

$$\int f d\mu = \int f^+ d\mu - \int f^- d\mu \,.$$

This is well-defined, since $f^+, f^- \leq |f|$, and by property (iv). One can then show properties (i)–(vi) hold for integrable functions.

References

Breiman, L. (1992). *Probability*. SIAM.

Dobrow, R. P. (2016). *Introduction to Stochastic Processes with R*. Wiley.

Durrett, R. (2005). *Probability: Theory and Examples*. Thomson, 3rd edition.

Grimmett, G. and Stirzaker, D. (2001). *Probability and Random Processes*. Oxford University Press.

Koralov, L. B. and Sinai, Y. G. (2010). *Theory of Probability and Random Processes*. Springer.

[4] almost everywhere, i.e. with probability 1