MATHEMATICAL METHODS I
MSC FINANCIAL ENGINEERING

RAYMOND BRUMMELHUIS

School of Economics, Mathematics and Statistics
Birkbeck College

Autumn Term 2004-2005

1. Introduction

Market prices of liquidly traded financial assets depend on a huge number of factors: macro-economic ones like interest rates, inflation, balanced or unbalanced budgets, micro-economic and business-specific factors like flexibility of labor markets, sale numbers, investments, and also on more elusive psychological factors like the aggregate expectations, illusions and disillusions of the various market players, both professional ones (stock brokers, market makers, fund managers, banks and institutional investors like pension funds), and humble private investors (like professors of finance, seeking to complement their modest revenue). Although many people have dreamed (and still do dream) of all-encompassing deterministic models for, for example, stock prices, the number of potentially influencing factors seems to high to realistically hope for such a thing. Already in the first half of the 20-th century researchers began to realize that a statistical approach to financial markets might be the best one. This started in fact right at the beginning of the century, in 1900, when a young French mathematician, Louis Bachelier, defended a thesis at the Sorbonne in Paris, France, on a probabilistic model of the French bourse. In his thesis, he developed the first mathematical model of what later came to be known as "Brownian motion", with the specific aim of giving a statistical description of the prices of financial transactions on the Paris stock market. The phrase with which he ended his thesis, that "the bourse, without knowing it, follows the laws of probability", are still a guiding principle of modern quantitative finance. Sadly, Bachelier's work was forgotten for about half a century, but was rediscovered in the '60's (in part independently), and adapted by the economist Paul Samuelson to give what is still the basic model of the price of a freely traded security, the so-called exponential (or geometric) Brownian motion[1]. The rôle of probability in finance has since then only increased,

[1]Peter Bernstein's book, Capital Ideas, gives a passionate account of the history of quantitative finance in the 20-th century

and quite sophisticated tools of modern mathematical probability, like stochastic differential calculus, martingales and stopping times, in combination with an array of equally sophisticated analytic and numerical methods, are routinely used in the daily business of pricing and hedging and risk assessment of ever more sophisticated financial products. The aim of the Mathematical methods Module of The MSC Financial Engineering is to teach you the necessary mathematical background to the modern theory of asset pricing. As such, it splits quite naturally into two parts: part I, to be taught during the Autumn semester, will concentrate on the necessary probability theory, while part II, to be given in the Spring Semester (Birkbeck College does not acknowledge the existence of Winter), will treat the numerical mathematics which is necessary to get reliable numbers out of the various mathematical models to which you will be exposed.

As already stated, Modern Quantitative Finance is founded upon the concept of stochastic processes as the basic description of the price of liquidly traded assets, in particular those which serve as underlying for derivative instruments like options, futures, swaps and the like. To price these derivatives it makes extensive use of what is called *stochastic calculus* (also known s *Ito calculus*, in honor of its inventor). Stochastic calculus can be thought of as the extension of ordinary differential calculus to the case where the variables (both dependent and independent) can be random, that is, have a value which depends on chance. Recall that in ordinary calculus, as invented by Newton and Leibniz in the 17-th century, one is interested in the behavior of functions of, in the simplest case, one independent variable,

$$y = f(x), \quad x \in \mathbb{R}.$$

In particular, it became important to know how the dependent variable y changes if x changes by some small amount Δx. The answer is given by the derivative, $f'(x)$, and by the relation:

$$\begin{aligned} \Delta y = \Delta f(x) &= f(x+\Delta x) - f(x) \\ &\simeq f'(x)\Delta x, \end{aligned}$$

where $\simeq$ means that we are neglecting higher powers of Δx. If we would not make such an approximation, we would have to include further terms of the Taylor series (provided f is sufficiently many times differentiable):

$$f(x+\Delta x) - f(x) = \\ f'(x)\Delta x + \frac{f''(x)}{2!}(\Delta x)^2 + \cdots + \frac{f^{(k)}(x)}{k!}(\Delta x)^k + \cdots \ .$$

It is convenient at this point to follow our 17-th and 18-th century mathematical ancestors, and introduce what are called *infinitesimals*

dx (also called *differentials*). These are non-zero quantities whose higher powers are equal to 0:

$$(dx)^2 = (dx)^3 = \cdots = 0.$$

We then simply write $f'(x) = (f(x+dx) - f(x))/dx$, or :

$$df(x) = f'(x)dx.$$

The mathematical problem with infinitesimals is that they cannot be real numbers, since no non-zero real number has its square equal to 0. For applied mathematics this is less of a problem, since simply thinks of dx as a number which is so small that its square and higher powers can safely be neglected. Physicists and engineers, unlike pure mathematicians, never stopped to use infinitesimals anyhow. Moreover, although mathematically not quite rigorous, if used with care, they generally lead to correct results, and most trained mathematicians can take any argument involving these infinitesimals and routinely convert it in a mathematically flawless proof leading to the same final result. What is more, infinitesimals can often be used to great effect to bring out the basic intuition which underlies results which might otherwise seem miraculous or obscure.

A case in point will be stochastic calculus, which aims to extend the notion of derivative to the case where x and y above are replaced by random variables (which we will systematically denote by capital letters, like X and Y). In this case the small changes dX will be stochastic also, and we have to try and establish a relation between an infinitesimal stochastic change of the independent (stochastic) variable , dX, and the corresponding change in $Y = f(X)$, $f(X+dX) - f(X)$. In fact, in most applications, such an X will be a specific instance $X = X_{t_0}$ of an infinite family of stochastic variables $(X_t)_{t\geq 0}$, where t is a positive real (non-stochastic!) parameter interpreted as time; for example, X_t might represent the market price of a stock at the future time $t \geq 0$. The first thing to do then is to establish a relation between dX_t and dt.

It turns out that, for a large class of stochastic processes called *diffusion processes* , or *Ito processes*[2], the right interpretation of dX_t is that of being a Gaussian random variable, with variance proportional to dt, and therefore basically of size $\simeq \sqrt{dt}$. This means that $(dX_t)^2$ will be of size $\simeq dt$, and can not be neglected anymore in the Taylor expansion of $f(X_t + dX_t)$. However, $(dX_t)^3 \simeq (dt)^{3/2}$ and this and higher powers will still count as 0. We therefore expect a formula like :

$$df(X_t) = f'(X_t)dX_t + \frac{1}{2}f''(X_t)(dX_t)^2,$$

which is basically the statement of the famous *Ito lemma*. Moreover, $(dX_t)^2$ turns out not to be stochastic anymore, but basically a multiple

[2]in honor of K. Itô , the inventor of stochastic calculus

of dt. The simplest stochastic process for which this program can be carried out is the afore-mentioned Brownian motion $(W_t)_{t\geq 0}$, which will in fact serve as the basic building block for more complicated processes. Brownian motion turns out to be *continuous*, in the sense that if two consecutive times $s < t$ are very close to each other, the (chance-) values W_s and W_t will be very close also, with probability 1. Modern Finance also uses other kinds of processes, which can suddenly *jump* from one value to another between one instant of time and the next. There is a similar basic building block in this case, which is called the *Poisson process*, in which the jumps are of a fixed size and occur at a fixed mean rate. In the much more complicated *Lévy processes*, which have recently become quite popular in financial modelling, the random variable can jump at different rates, and the jump sizes will be stochastic also, instead of being fixed. For all of these processes, people have established analogues of the Ito lemma mentioned above. Although we will briefly look at these more complicated processes at the end (time allowing), our emphasis will be on Brownian motion and diffusion processes.

To properly set up all of this in a mathematically rigorous way one is usually obliged to undergo (one is tempted to say, "suffer") an extensive preparation in abstract probability theory from what is called the measure theoretic point of view. It turns out, however, that the basic rules of Ito calculus can be explained, and motivated, using a more intuitive, 19-th century, approach to probability, if one is willing to accept infinitesimals. In the first part of these lectures we will follow such an approach, with the aim of familiarizing you as quickly as possible with stochastic calculus, including processes with jumps. This will probably take the first 4 to 5 weeks. Afterwards we will take a closer look at the measure-theoretic foundations of modern probability, and at least sketch how the material explained in the first half fits into the new, more abstract framework, and can be used to make things rigorous. I would like to stress, however, that mathematical rigor is not the only aim there. An equally important point is that the measure theoretic approach to probability will allow us to formalize the concepts like that of "information contained in a rv or in a stochastic process up till time t", "martingale" and of "stopping times". A martingale is a stochastic process which can be thought of as a fair (gambling) game, in the sense that at each point in time and given all information on how the game as developed up till that point in time, one's expected gain when continuing to play still equals one's expected loss (think of repeatedly tossing a perfect coin). In a world without interest rates, idealized stock prices should be martingales: this is one way of formulating the so-called Efficient Market Hypothesis. Stopping times are (future) random times at which you (or somebody else) will have taken some action depending on the information that will have been made

available at that future time. These are basic for, for example, understanding American style contracts, or any kind of situation in which investors are free to chose their time of action. Finally, the abstract approach will allow us to "change probabilities", and clarify the concept of risk neutral investors as (hypothetical) investors which accord different probabilities to the same events as non-risk neutral ones, to the effect that they do not require to be awarded for risk taking. All the material to be presented in the second half of these lectures will be fundamental for the Spring term of the Pricing module, where we will explain pricing using the martingale method.

The end of an example, exercise, definition, theorem, lemma or proposition will be indicated by a double Dollar sign, $$. Starred remarks, examples, etc. mostly serve to put the material in a wider mathematical context, and can be skipped without loss of continuity.

2. Review of Probability Theory (19-th Century Style)

2.1. **Real random variables.** In the first half of these lecturers we will use an informal approach to probability theory, using "probability" and "random variables" as the, for the moment unexplained, primitive notions of the theory (much like "points" and "lines" in Geometry), for which we will trust upon commonly shared intuition. In particular, a *real-valued random variable* will be a quantity, whose exact value we are not sure about, but of which we do know the various probabilities that it will lie in any given interval of real numbers. That is, a real random variable X will (at this stage of the theory) be characterized by various probabilities, like

$$\mathbb{P}(a \leq X \leq b) := (\text{Probability that } X \text{ will lie between } a \text{ and } b\ ), \tag{1}$$

which, by definition, will be a number between 0 and 1:

$$\mathbb{P}(a \leq X \leq b) \in [0,1].$$

Here $\mathbb{P}$ stands for 'probability', and a and b can be any pair of real numbers. We also allow a or b to be $\mp\infty$, respectively. By convention, $X < \infty$ and $X \leq \infty$ are trivially fulfilled statements, whose probability is 1, and $X < -\infty$ is an empty statement, whose probability is 0. We obviously want $\mathbb{P}(a \leq X \leq b)$ to be a number between 0 and 1.

Random variables will systematically be denoted by capital letters X, Y, Z, etc., while ordinary real numbers will be denoted by lower case letters x, y, z (this convention will later on be extended to vector valued random variables and ordinary vectors). We will usually abbreviate 'random variable' by 'rv'.

Easiest to understand are probably what are called *discrete random variables*, those which effectively will only assume values in some discrete (possibly infinite) set of real numbers $\{x_1, x_2, \cdots, \}$. Such a discrete rv X is completely determined by the various probabilities that X actually equals one of these x_j:

$$p_j = \mathbb{P}(X = x_j). \tag{2}$$

This clearly implies that

$$\mathbb{P}(a \leq X \leq b) = \sum_{j:\ a \leq x_j \leq b} p_j.$$

In particular, this probability is 0 if none of the x_j's lie between a and b.

The following is an example of a discrete rv which is basic in mathematical finance:

Example 2.1. In the *binomial option pricing model* one supposes that S_N, the price of the underlying security N days into the future, can

take on the values

$$d^N S_0, d^{N-1} u S_0, \cdots d^{N-j} u^j S_0, \cdots, u^N S_0 \quad (0 \leq j \leq N),$$

S_0 being today's price and u and d two fixed positive real numbers (we are assuming that in any one day, the stock's price can only go up or down by a fixed fraction, u respectively d). The probability that S_N is any of these values then is defined as:

$$\mathbb{P}\left(S_N = d^{N-j} u^j S_0\right) = \left(\begin{array}{c} N \\ j \end{array} \right) p^{N-j} (1-p)^j.$$

Here p is the probability of a daily 'up move' (price moving from S to uS in going from one day to the next), and $1-p$ that of a 'down move' ($S \to dS$). $$

Exercise 2.2. To have a well-defined discrete random variable, one needs that

$$\sum_j p_j = 1.$$

(Why?) Check that this is indeed the case for the binomial model defined just now.
(*Hint:* Use a well-known formula from Algebra which is also called 'binomial'.) $$

Another important example of a discrete rv is a Poisson rv:

Example 2.3. A *Poisson rv* N is a discrete rv, taking its values in $\mathbb{N} = \{0, 1, 2, \cdots\}$, for which

$$\mathbb{P}(N = k) = p_k = \frac{\lambda^k}{k!} e^{-\lambda}.$$

Here $\lambda > 0$ is a parameter. Here $\sum_{k=1}^{\infty} p_k = 1$ since $e^{\lambda} = \displaystyle\sum_{k=0}^{\infty} \frac{\lambda^k}{k!}$. $$

One can go a long way using only discrete rv, but at some point it becomes extremely convenient[3] to dispose of what are called *continuous random variables*. These do not take on any particular real value with a non-zero probability, but their probable values are, so to speak, spread out over entire intervals, and often even over the whole of $\mathbb{R}$. For such a rv X we will have that $\mathbb{P}(X = a) = 0$ for any $a \in \mathbb{R}$, but typically $\mathbb{P}(a - \varepsilon \leq X \leq a + \varepsilon) \neq 0$, for any $\varepsilon > 0$. An very important example of such a rv is what is called a standard normal random variable:

Example 2.4. X is called a *standard normal* random variable (one also often uses the term 'Gaussian rv') if, for any $a < b$,

$$\mathbb{P}(a < X \leq b) = \int_a^b e^{-x^2/2} \frac{dx}{\sqrt{2\pi}}.$$

[3] and even essential: see the Central limit Theorem below!

$$

The condition that $\mathbb{P}(-\infty < X < \infty)$ follows from the classical integral:

$$\int_{-\infty}^{\infty} e^{-x^2/2} dx = \sqrt{2\pi};$$

see any reasonable calculus textbook.

Standard normal rvs are a sort of 'universal' random variables in probabilistic modelling, for reasons which will become clair when we will discuss the Central Limit Theorem below.

An economical way of specifying all probabilities associated to a rv is by defining what is called the cumulative distribution function:

Definition 2.5. The *cumulative distribution function or cdf* (often also called the *probability distribution function*, or simple the distribution function) of a rv X is the function $F_X : \mathbb{R} \to [0, 1]$ defined by:

$$F_X(x) = \mathbb{P}(X \leq x). \tag{3}$$

$$

Example 2.6. The cdf of a discrete rv defined by (2) is:

$$F_X(x) = \sum_{j: p_j \leq x} p_j.$$

Observe that F_X jumps by an amount of p_j at the points a_j (draw a graph!). $$

From F_X we can find the other probabilities, like for example

$$\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a).$$

We list the properties which characterize a cdf F_X:

- $F_X(x) \to 0$ as $x \to -\infty$ and $F_X(x) \to 1$ as $x \to \infty$(since the probability that X will take *some* real value is 1).
- F_X is *increasing:* if $x_1 \leq x_2$ then $F_X(x_1) \leq F_X(x_2)$ (since $X \leq x_1$ will imply $X \leq x_2$).
- F_X is what is called *right continuous*: $\lim_{\varepsilon > 0, \varepsilon \to 0} F_X(x + \varepsilon) = F_X(x)$, for all $x \in \mathbb{R}$.

*__Remark 2.7.__ The mathematical reason for this third property is perhaps not yet very clear at this point; the right continuity is in fact connected with the fact that we have a $\leq$-sign in (3); if we would have defined F_X with a $<$-sign, we would have had left-continuity. For the moment one can just consider this third condition as a rule for what to do in points where F_X jumps. We note in this connection that it is a general mathematical result that an increasing function like F_X can only have jump discontinuities. $$

Conversely, any function $F : \mathbb{R} \to [0, 1]$ with the above three properties will define a rv X by taking F as its cdf, that is, by specifying that

$$\mathbb{P}(X \leq x) = F(x),$$

by definition. Equivalently, $F_X = F$.

An important class of rv is those having a probability density function:

Definition 2.8. A rv X is said to have a probability density function of pdf if its cdf is of the form:

$$F_X(x) = \int_{-\infty}^{x} f(x)dx,$$

for some (integrable) function $f : \mathbb{R} \to \mathbb{R}$. This function f is (essentially[4]) unique, and we write $f = f_X$, to stress the dependence on X.

$$$$

If X has a pdf, F_X is continuous, and for reasonable (say, continuous) f, F_X will also be differentiable, with derivative

$$F'_X(x) = f(x).$$

A standard normal variable has pdf

$$\frac{e^{-x^2/2}}{\sqrt{2\pi}}.$$

Discrete random variables, such as Poisson rvs, do not have a pdf: since their cdf have jumps, and are therefore not continuous.

*__Remark 2.9.__ One can construct very curious cdf's which are continuous everywhere, have derivatives in *almost all*[5] their points, but which do *not* have a pdf. If this derivative is equal to 0 almost everywhere, such a cdf (and its associated rv) is called totally singular. Note that the condition of being continuous prevents such a cdf to have jumps; in particular, $\mathbb{P}(X = a)$ will still be 0, for any $a \in \mathbb{R}$. and we are still far from having the cdf of a discrete rv. $$

A function $f = f(x)$ will be the pdf of a random variable X iff[6] the function $F(x) = \int_{\infty}^{x} f(y)dy$ has the properties of a cdf. This leads to the following characterizing properties of a pdf:

- $f(x) \geq 0$ everywhere (corresponding to F_X being increasing).

[4] we might for example change f in a finite number of points without changing the integral

[5] a term which has to be understood in a certain technical sense, which we'll explain when discussing measure theoretic probability

[6] a very usefull abbreviation, standing for'if and only if'

- $\int_{-\infty}^{\infty} f(x)dx = 1$. (corresponding to $F_X(x) \to 1$ as $x \to \infty$ or, equivalently, to the total probability having to sum to 1).

Continuity, and therefore right continuity, is automatic for functions $F(x)$ which can be written as integrals.

Random variables X having a pdf are the easiest to work with. Although the probability that such an X will take on precisely the value $x \in \mathbb{R}$ is equal to 0 for any real x, there is a useful alternative. Let dx be an (calculus-style) infinitesimal:

$$dx \neq 0,\ (dx)^2 = (dx)^3 = \cdots = 0.$$

(Such infinitesimals do not really exist, but we think of them as numbers which are so small, that their squares may be safely neglected in any computation. An operative definition, in a computing context, might be to take dx so small that all the significant digits of its square are equal to 0, within machine precision or within the precision which is significant for the problem at hand.) We then think of $f_X(x)dx$ as being the probability that X lies in the infinitesimally small interval between x and $x + dx$:

$$\mathbb{P}(X \in [x, x + dx]) = f_X(x)dx.$$

Often we will be quite sloppy in our notations, and simply write

$$\mathbb{P}(X = x) = f_X(x),$$

although, strictly speaking, the left hand side is 0 here ... .

To see how this works, consider the definition of the *mean* of a continuous rv X. The mean of a discrete rv is equal to the sum of the possible values it can assume times the probability that it will take on that value:

$$\sum_j a_j \mathbb{P}(X = a_j).$$

For a rv X having a pdf one would like to take the sum over all possible x of $x \cdot \mathbb{P}(X \in [x, x + dx])$. The continuous analogue of a sum being an integral, this leads to the following definition: :

$$\mathbb{E}(X) = \int_{\mathbb{R}} x f(x)dx. \tag{4}$$

More generally, and for similar reasons, when we consider functions $g(X)$ of such a random variable X, its mean is given by the *very important* formula:

$$\mathbb{E}(g(X)) = \int_{\mathbb{R}} g(x)f(x)dx, \tag{5}$$

provided this integral has a sense and is finite. The formula is very important because, typically in Finance, present prices, e.g. of options, can be expressed as means and in 99% (or perhaps even 100 %) of the

cases when you will be analytically evaluating such a price, you will have to use (5).

Particular examples of (5) are of course the mean $\mathbb{E}(X)$ of X itself (corresponding to $g(x) = x$) and, putting $\mu_X = \mathbb{XE}(X)$, the *variance* of X:

$$\text{Var}(X) = \mathbb{E}\left((X - \mu_X)^2\right) = \int_{\mathbf{R}} (x - \mu_X)^2 f(x)dx, \tag{6}$$

corresponding to $g(x) = (x - \mu_X)^2$ and again assuming the integral is finite. We often write

$$\text{Var}(X) = \sigma_X^2,$$

where $\sigma_X = \sqrt{\text{Var}(X)}$ is the *standard deviation*; σ_X is a measure of how much, in the mean, X differs from its mean, μ_X[7].

A very useful computational rule is that

$$\text{Var}(X) = \mathbb{E}(X^2 - (\mathbb{E}(X))^2 \quad \text{(exercise!)}$$

Higher moments are often also very useful in finance, in particular the following two:

$$\text{Skewness:} \quad s(X) = \mathbb{E}\left(\frac{(X-\mu)^3}{\sigma^3}\right) = \int_{\mathbb{R}} \left(\frac{x-\mu}{\sigma}\right)^3 f(x)dx \tag{7}$$

$$\text{Kurtosis:} \quad \kappa(X) = \mathbb{E}\left(\frac{(X-\mu)^4}{\sigma^4}\right) = \int_{\mathbb{R}} \left(\frac{x-\mu}{\sigma}\right)^4 f(x)dx \tag{8}$$

Skewness is an indication of whether the pdf is tilted to the right or to the left of it's mean: if $s(X) > 0$, then X has a bigger probability to have values bigger μ than to have values smaller than μ, and vice versa. A big kurtosis is an indication that $|X|$ can have large values with relatively big probability. These quantities play an important rôle in the econometric analysis of financial returns. The following example computes them for the benchmark case of a normal rv with arbitrary mean and variance.

Example 2.10. *(General normal or Gaussian variables)* A rv X is said to be normally distributed with mean μ and variance σ^2 if X has probability density:

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}. \tag{9}$$

In this case we write

$$X \sim N(\mu, \sigma^2).$$

[7]another measure of this deviation could be something like $\mathbb{E}(|X - \mu_X|)$, but experience as learned that quadratic expressions like variances are much easier to compute with

The standard normal rv corresponds to $\mu = 0$ and $\sigma = 1$. One easily checks that this is a correct definition, since this function is non-negative, and since its total probability is equal to 1:

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} dx = 1.$$

(The easiest way to see this is by making the successive changes of variables $x \to x + \mu$ and $x \to \sigma x$ to reduce to the case of a standard normal variable.)

If $x \sim N(\mu, \sigma^2)$, then it's mean, variance, skewness and kurtosis are, respectively:

$$\begin{aligned}
&\text{mean: } \int x e^{-(x-\mu)^2/2\sigma^2} \frac{dx}{\sqrt{2\pi\sigma^2}} = \mu, \\
&\text{variance } \int (x-\mu)^2 e^{-(x-\mu)^2/2\sigma^2} \frac{dx}{\sqrt{2\pi\sigma^2}} = \sigma^2, \\
&\text{skewness } \frac{1}{\sigma^3} \int (x-\mu)^3 e^{-(x-\mu)^2/2\sigma^2} \frac{dx}{\sqrt{2\pi\sigma^2}} = 0 \text{ (by symmetry)} \\
&\text{kurtosis } \frac{1}{\sigma^4} \int (x-\mu)^4 e^{-(x-\mu)^2/2\sigma^2} \frac{dx}{\sqrt{2\pi\sigma^2}} = 3.
\end{aligned}$$

(To do these integrals, first make a change of variables, as above, to get rid of the μ and the σ.) $$

If X is any, not necessarily normally distributed, rv, one often compares its kurtosis with that of a normal distributed rv having the same variance. This leads to the concept of *excess kurtosis*:

$$\kappa_{\text{exc}}\ (X) = \kappa(X) - 3. \tag{10}$$

If the excess kurtosis is positive, the pdf of X is interpreted to have more probability mass in the tails, or *fatter tails* , than that of a comparable normal distribution, with the same mean and variance as X.

*__Example 2.11.__ Random variables do not need to have a well-defined mean or variance: the following is a classical example, dating back to the 1800's, when it caused much controversy among the French probabilists. A random variable X is said to be *Cauchy*, or have a Cauchy distribution, if its pdf is given by:

$$\frac{1}{\pi} \frac{1}{1+x^2}.$$

This gives rise to a well-defined rv, since

$$\frac{1}{\pi} \int_{-\infty}^{\infty} \frac{dx}{1+x^2} = 1,$$

as one easily checks, using that a primitive of $(1+x^2)^{-1}$ is $\arctan x$. Is the mean of X well-defined? This is not quite clear: on the one hand

one might argue that (forgetting momentarily about the $1/\pi$ in front):

$$\int_{-\infty}^{\infty} \frac{x}{1+x^2} dx = \lim_{R\to\infty} \int_{-R}^{R} \frac{x}{1+x^2} dx = 0,$$

by symmetry. On the other hand, if one takes the limit of the integrals over asymmetric intervals expanding to the whole of $\mathbb{R}$, the answer comes out quite differently. For example:

$$\begin{aligned} \lim_{R\to\infty} \int_{-R}^{2R} \frac{x}{1+x^2} dx &= \lim_{R\to\infty} [\frac{1}{2}\log(1+x^2)]_{-R}^{2R} \\ &= \lim_{R\to\infty} \frac{1}{2}\log\left(\frac{1+4R^2}{1+R^2}\right) \\ &= \log 2 \neq 0! \end{aligned}$$

(Here log stands for the natural logarithm, with basis e.)
What is going on here? Mathematically speaking,

$$\int_{-\infty}^{\infty} f(x)dx = \lim_{a\to-\infty, b\to\infty} \int_{a}^{b} f(x)dx$$

will be well-defined, that is, will be independent of the way a and b tend to $\mp\infty$, if

$$\lim_{R\to\infty} \int_{-R}^{R} |f(x)|dx < \infty,$$

(readers familiar with the Lebesgue integral will note that this is equivalent with the Lebesgue integrability of $|f|$ on $\mathbb{R}$), and it is clear that in our example,

$$\lim_{R\to\infty} \int_{R}^{R} \frac{|x|}{1+x^2} = \lim_{R\to\infty} \log(1+R^2) = \infty.$$

Even if one argues that the symmetric definition of the mean is natural, since in our example the pdf is symmetric, and therefore puts $\mu_X = \mathbb{E}(X) = 0$, one runs into problems with the variance, since it is clear that, e.g.

$$\lim_{R\to\infty} \int_{-R}^{R} \frac{x^2}{1+x^2} dx = \infty.$$

(Use integration by parts, or estimate the integral from below by, for example, $\int_1^R x^2/(1+x^2)dx \geq \int_1^R (1/2)dx = (R-1)/2 \to \infty$.)

The Cauchy distribution is a particular example of a more general class of distributions called the *Lévy stable distributions*, which also include the normal distribution (which actually is the only member of this class having a finite variance), and to which we will (hopefully) devote some time at the end of these lectures. Lévy stable distributions have been proposed as more accurate models of financial asset returns than the traditional normal distributions (following pioneering work by B. Mandelbroit and E. Fama in the '60's), but are more difficult to

work with for a number of technical reasons, not the least of which is their infinite variance. $$

*__Remark 2.12.__ How to define the mean $\mathbb{E}(X)$ and, more generally $\mathbb{E}(g(X))$ if X is neither discrete nor has a probability density? This is not quite so obvious, but it turns out that for reasonable functions $g = g(x) : \mathbb{R} \to \mathbb{R}$ the following definition is a natural generalization of (5):

$$\mathbb{E}(g(X)) = \lim_{N\to\infty} \sum_{j=-\infty}^{\infty} g\left(\frac{j}{N}\right)\left(F_X\left(\frac{j+1}{N}\right) - F_X\left(\frac{j}{N}\right)\right). \tag{11}$$

This formula is motivated by the classical construction of an integral $\int_a^b g(x)dx$ as limit of sums over rectangles filling in the surface under the graph of g, as you may recall from your calculus course (indeed, the latter corresponds formally to taking $F_X(x) = x$, although this is not a pdf). We will denote the right hand side of 11) by :

$$\int_{\mathbb{R}} g(x)dF_X(x), \tag{12}$$

with the understanding that if F_X is differentiable (and X has a pdf $F_X' = f_X$), then

$$dF_X(x) = f_X(x)dx,$$

so that we get (5) again. $$

*__Exercise 2.13.__ Convince yourself that, if X is a discrete rv, taking values in $\{a_1, a_2, \cdots\}$ with probabilities $p_1, p_2, \cdots$, then for continuous g,

$$\int_{\mathbb{R}} g(x)dF_X = \sum_j p_j g(a_j). \tag{13}$$

As special cases we re-obtain the classical formulas for the mean and variance f a discrete rv:

$$\mathbb{E}(X) = \mu_X = \sum_{j=1}^{n} p_j a_j,$$

and

$$\mathrm{Var}(X) = \sum_{j=1}^{n} (a_j - \mu_X)^2 p_j.$$

What about (13) when g has a jump in $x = a_1$ and is right-continuous there? And what if it is left-continuous? (Take $a_1 = 0$, to simplify). $$

__Exercise 2.14.__ Compute the mean and variance of a Poisson random variable. $$

Exercise 2.15. Let X be a rv with pdf $f = f_X$. Show that X^2 also has a pdf, which is given by:

$$\frac{1}{2\sqrt{x}}\left(f(\sqrt{x}+f(-\sqrt{x})\right).$$

As an application, compute the pdf of X^2 when X is standard normal. The result is called a $\chi^2_{(1)}$ or χ^2-*distribution with one degree of freedom.* \$\$

Exercise 2.16. Let Z be a standard normal variable. Compute the pdf of $X = e^Z$; this is called an log-normal distribution. Compute the mean and variance of X. \$\$

Exercise 2.17. A *Student t-distribution with $\nu > 2$ degrees of freedom* has a pdf of the form:

$$t_\nu(x) = C_\nu \left(1 + \frac{x^2}{\nu - 2}\right)^{-\nu/2},$$

C_ν being a normalization constant which is put there to insure that $\int t_\nu(x)dx = 1$. Show that if X is Student, then its mean exists. Also show that its variance exists iff $\nu > 3$, and its skewness and kurtosis iff, respectively, $\nu > 4$, $\nu > 5$.

(*Hint:* Use that the integral $\displaystyle\int_1^\infty dx/x^\alpha$ is finite iff $\alpha > 1$.) \$\$

2.2. **Random Vectors and Families of Random Variables.** Consider a vector of real random variables $(X_1, \cdots, X_N)$. When can we say to know such a random vector? Clearly, we must know the probability distribution F_{X_j} of each of the X_j individually, but we need to know more. For example, we would also want to know joint probabilities such as the one that $a_1 < X_1 \leq b_1$ while $a_2 < X_2 \leq b_2$, and similarly for three or more of the X_j's. In fact, all this information can be gotten out of the *joint distribution function*,

$$F_{X_1,\cdots,X_N}(x_1, \cdots, x_N) := \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, \cdots X_N \leq x_N), \tag{14}$$

the probability that, simultaneously, $X_1 \leq x_1$ *and* $X_2 \leq x_2$, etc. One can show that one can get joint probabilities like:

$$\mathbb{P}(a_1 < X_1 \leq b_1, a_2 < X_2 \leq b_2, \cdots, a_N \leq X_N \leq b_N), \tag{15}$$

from (14), by algebraic manipulations, but we will leave this as an exercise to the interested reader (the answer can be found in most books on probability theory; try to work it out for, two variables (X_1, X_2)).

We will again mostly work with $(X_1, \cdots, X_N)$'s having a multivariate probability density, in the (obvious) sense that there is a function $f_{X_1,\cdots,X_N} : \mathbb{R}^N \to \mathbb{R}_{\geq 0}$ such that

$$F_{X_1,\cdots,X_N}(x_1, \cdots, x_N) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_N} f_{X_1,\cdots,X_N}(y_1, \cdots, y_N)dy_1 \cdots dy_N. \tag{16}$$

We then say that $(X_1, \cdots, X_N)$ has joint pdf $f_{X_1,\cdots,X_N}$. Note that in this case (15) simply equals

$$\int_{a_1}^{b_1} \cdots \int_{a_N}^{b_N} f_{X_1,\cdots,X_N} dy_1 \cdots dy_N,$$

where, to simplify the formulas, we will often leave out the variables of $f_{X_1,\cdots,X_N}$. Note that the definition of a joint pdf implies that

$$f_{X_1,\cdots,X_N)}(x_1, \cdots, x_N) = \frac{\partial^N}{\partial x_1 \cdots \partial x_N} F_{X_1,\cdots,X_N}(x_1, \cdots, x_N).$$

If $(X_1, \cdots, X_N)$ has joint pdf $f_{X_1,\cdots,X_N}$, then the natural definition of the expectation of a function $g(X_1, \cdots, X_N)$ of the $X_1, \cdots, X_N$ is:

$$\mathbb{E}(g(X_1, \cdots, X_N)) \tag{17}$$
$$:= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} g(x_1, \cdots, x_N) f_{X_1,\cdots,X_N}(x_1, \cdots, x_N) dx_1 \cdots x_N.$$

We will usually write this more briefly as:

$$\mathbb{E}(g(X_1, \cdots, X_N)) = \int_{\mathbb{R}^N} g f_{X_1,\cdots,X_N} dx,$$

with $dx = dx_1 \cdots dx_N$.

In particular, we define the means μ_{X_j} and the *covariances* $\mathrm{Cov}(X_i, X_j)$ by:

$$\mu_{X_j} = \mathbb{E}(X_j) = \int_{\mathbb{R}^N} \int x_j f_{X_1,\cdots,X_N} dx, \tag{18}$$

and

$$\begin{aligned} \mathrm{Cov}(X_i, X_j) &= \mathbb{E}\left((X_i - \mu_{X_i})(X_j - \mu_{X_j})\right) \\ &= \int_{\mathbb{R}^N} (x_i - \mu_{X_i})(x_j - \mu_{X_j}) f_{X_1,\cdots,X_N} dx. \end{aligned} \tag{19}$$

The following is the multi-variable generalization of a normal rv:

Example 2.18. *(jointly normally distributed random vectors)* Let $\mathbb{V} = (V_{ij})_{1 \le i,j \le N}$ be a non-singular symmetric $N \times N$-matrix:

$$V_{ij} = V_{ji} \in \mathbb{R}, \quad \det(\mathbb{V}) \neq 0.$$

We say that $(X_1, \cdots, X_N)$ are *jointly normally distributed with mean* $\mu = (\mu_1, \cdots, \mu_N)$ *and variance-covariance matrix* $\mathbb{V}$ if their joint pdf is equal to:

$$f_{X_1,\cdots,X_N}(x_1, \cdots, x_N) = \frac{\exp(-\langle x, \mathbb{V}^{-1} x \rangle / 2}{(2\pi)^{N/2} \sqrt{\det(\mathbb{V})}}. \tag{20}$$

Here $\mathbb{V}^{-1}x$ stands for the inverse of $\mathbb{V}$ applied to $x = (x_1, \cdots x_N) \in \mathbb{R}^N$, and $\langle \cdot, \cdot \rangle$ stands for the Euclidian inner product on $\mathbb{R}^N$:

$$\langle x, y \rangle = x_1 y_1 + \cdots + x_N y_N.$$

(Often also written as $x^t y$, where ^{t} stands for 'transpose'). $$

Remark 2.19. One can check that if

$$(X_1, \cdots, X_N) \sim N(\mu, \mathbb{V}),$$

then

$$\mathbb{E}(X_j) = \mu_j,$$

and

$$\mathrm{Cov}(X_i, X_j) = V_{ij}.$$

This can be verified directly by using a bit of linear algebra and multi-variable calculus, by diagonalizing $\mathbb{V}$ using a suitable rotation of $\mathbb{R}^N$, and using the change of variables formula for multiple integrals, details left to the interested reader. An easier and more natural way to deal with multi-variate normals will be introduced in section 3.4 below. $$

From the joint distribution of $(X_1, \cdots, X_N)$ we can re-construct the single cdf's of the X_j by taking what are called the *marginals* of $F_{X_1,\cdots,X_N}$. For example, since, trivially, any $X_j < \infty$, we can write:

$$\begin{aligned} F_{X_1}(x_1) &= \mathbb{P}(X_1 \leq x_1) \\ &= \mathbb{P}(X_1 \leq x_1, X_2 < \infty, \cdots, X_N < \infty) \\ &= F_{X_1,\cdots,X_N}(x_1, \infty, \cdots, \infty), \end{aligned}$$

the second equality being true since the condition $X_j < \infty$ is always trivially satisfied, and where we've put:

$$F_{X_1,\cdots,X_N}(x_1, \infty, \cdots, \infty) = \lim_{y_2 \to \infty} \cdots \lim_{y_N \to \infty} F_{X_1, y_2 \cdots, y_N}(x_1, y_2 \cdots, y_N).$$

This is called taking a marginal of the joint distribution $F_{X_1,\cdots,X_N}$. In the case that $(X_1, \cdots, X_N)$ has a joint pdf,

$$F_{X_1}(x_1) = \int_{-\infty}^{x_1} \int_{-\infty}^{\infty} \cdots \int_{\infty}^{\infty} f_{X_1,\cdots,X_N}(y_1, \cdots, y_N) dy_2 \cdots dy_N.$$

We can in particular differentiate w.r.t. x_1, and find that X_1 also has a pdf, given by:

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} \cdots \int_{\infty}^{\infty} f X_1, \cdots, X_N(x_1, y_2, \cdots, y_N) dy_2 \cdots dy_N.$$

That is, we find the pdf of X_1 by integrating out all variables other than x_1. The analogous construction applies for any f_{X_j}.

*__Remark 2.20.__ We will forgo the general definition of $\mathbb{E}(g(X_1,\cdots,X_N)$ when there is no density: we will at most need this when $X_1\cdots,X_N$ are independent (see the following section), in which case $dF_{X_1,\cdots,X_N}$ can be interpreted as a repeated integral with respect to $dF_{X_1}(x_1)\cdots dF_{X_N}(x_N)$. We will return to this point when (and if) needed. $$

We will very soon need to go beyond finite vectors of random variables, and consider infinite families of these. Indeed, a *continuous time stochastic process* is defined as a collection of random variables $(X_t)_{t\geq 0}$, one for each positive t, the latter playing the role of time. How do we specify such a stochastic process? This turns out to be a bit delicate, in particular as to the question of when to identify two such stochastic processes, but for the moment we will use the following working definition:

Definition 2.21. *(Stochastic processes, provisional working definition)* A continuous-time stochastic process is a collection of random variables, X_t, one for each $t\geq 0$ such that, for any finite collection of times $\{t_1,t_2,\cdots,t_N\}$ we know the joint probability distribution $F_{X_{t_1},\cdots,X_{t_N}}:\mathbb{R}^N\rightarrow[0,1]$ of $(X_{t_1},\cdots,X_{t_N})$ (N can be arbitrarily big, here). $$

*__Remark 2.22.__ The delicacy here resides in the fact that t ranges over a continuous set. *Discrete time stochastic processes* $(X_n)_{n\in\mathbb{N}}$ are less problematic in the sense that these are completely determined by all joint distributions $F_{X_1,\cdots,X_N}$, for arbitrarily large N.

For continuous t one usually includes a (left- or right) continuity condition on what are called the *sample trajectories* $t\rightarrow X_t$. To properly define these latter we will need to turn to the measure theoretic approach to probability in the second half of these lectures. $$

2.3. **Independent Random Variables and Conditional Probabilities.** The concept of independent rv is basic in probability and statistics. Let us consider a pair of random variables (X,Y) with joint probability distribution $F_{X,Y}$.

Definition 2.23. *(independent random variables.)* Two rv X and Y are *independent* if, for all x,y,

$$F_{X,Y}(x,y)=F_X(x)F_Y(y) \tag{21}$$

$$

One easily checks that if (X,Y) has a joint pdf, then X and Y are independent iff

$$f_{X,Y}(x,y)=f_X(x)f_Y(y), \tag{22}$$

where f_X and f_Y are the marginal pdfs:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy, \quad \text{etc.}$$

One easily shows from this that $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ (the double integral just becomes a product of two one-dimensional integrals). More generally (and for the same reason) we have:

Proposition 2.24.

$$X, Y \text{ independent } \Rightarrow \mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y)). \tag{23}$$

for any two functions $g = g(x)$ and $h = h(y)$ for which these expectations are well-defined.

*__Remark 2.25.__ Having the right hand side of (23) for a sufficiently large class of functions g and h, for example for all bounded continuous functions, is in fact equivalent to definition (2.23). $$

If we recall that covariance of two rv X and Y:

$$\operatorname{cov}(X,Y) = \mathbb{E}\left((X-\mu_X)(Y-\mu_Y)\right),$$

where $\mu_X = \mathbb{E}(X)$ and $\mu_Y = \mathbb{E}(Y)$ are the means, then (23) implies:

$$X, Y \text{ independent } \Rightarrow \operatorname{cov}(X,Y) = 0.$$

For jointly normal rvs there is a well-known converse, which we mention without proof (the easiest proof uses the concept of *characteristic function* of a random vector):

Proposition 2.26. *If (X,Y) is jointly normally distributed, then $cov(X,Y) = 0$ implies that X and Y are independent. $$*

But **ATTENTION:** 2.26 is **NOT** true if X and Y are **NOT** jointly normal: "having covariance 0" is in general much weaker than being independent, as the following example shows:

Example 2.27. Let $X \sim N(0,1)$ be a standard normal rv, and let $Y = X^2 - 1$. Then $\mathbb{E}(X) = \mathbb{E}(Y) = 0$ (the latter since $\mathbb{E}(X^2) = 1$, and

$$\begin{aligned} \operatorname{cov}(X,Y) &= \mathbb{E}(XY) \\ &= \mathbb{E}\left(X(X^2-1)\right) \\ &= \mathbb{E}(X^3) - \mathbb{E}(X) \\ &= 0, \end{aligned}$$

since $\mathbb{E}(X) = \mathbb{E}(X^3) = 0$. So X and Y have 0 covariance. However, they are not independent: intuitively this is clear, since Y is even a function of X, and therefore as dependent on X as can be! Formally, if we take $g(x) = x^2 - 1$ and $(x) = x$, then

$$\mathbb{E}\left(g(X)h(Y)\right) = \mathbb{E}\left((X^2-1)^2\right) > 0,$$

contradicting independence, by proposition 2.24. $$

Working a little harder, one can find a joint pdf $f_{X,Y}$ such that $\mathrm{Cov}(X,Y)=0$, and such that both its marginals are normal, but such that X and Y are still *not* independent. This shows that one has to be very careful on how to formulate (2.26): one cannot replace "(X,Y) normally distributed" by "X and Y normally distributed".

If we recall, for later use, the definition of the correlation coefficient:

$$\rho(X,Y)=\frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(X)}}, \tag{24}$$

which is always a number between -1 and 1, then X and Y independent implies that $\rho(X,Y)=0$, but the converse is not true, except again for Gaussian rv.

The previous considerations generalize naturally to N-tuples of random variables: $X_1,\cdots,X_N$ will be called independent iff

$$F_{X_1,\cdots,X_N}(x_1,\cdots,x_N)=F_{X_1}(x_1)\cdots F_{X_N}(x_N), \tag{25}$$

for any $x_1,\cdots,x_N)\in\mathbb{R}^N$, and we have the natural generalization of (23), which we will leave as an exercise.

We next turn to the concept of conditional probability for pairs of random variables. The discussion here will be limited to rvs having densities. The general case needs a considerably more abstract approach, and will be treated in the second half of these lectures.

Recall, form elementary probability theory, that the conditional probability of some event A happening, given that B has (or will have) happened, is defined as

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

For X,Y two rv with probability densities f_X and f_Y, one can therefore compute the *conditional probability of X being in $[x,x+dx]$ given that Y is in $[y,y+dy]$* as:

$$\begin{aligned}
&\mathbb{P}(X\in(x,x+dx)|Y\in(y,y+dy))\\
=\ &\frac{\mathbb{P}(X\in(x,x+dx),y\in(y,y+dy))}{\mathbb{P}(Y\in(y,y+dy))}\\
=\ &\frac{f_{X,Y}(x,y)dxdy}{f_Y(y)dy}\\
=\ &\frac{f_{X,Y}(x,y)}{f_Y(y)}dx,\ )
\end{aligned} \tag{26}$$

assuming of course that $F_Y(y)\neq 0$. We will often simply write this as:

$$\mathbb{P}(X=x|Y=y)=\frac{f_{X,Y}(x,y)}{f_Y(y)},$$

forgetting about the dx, and read the left hand side as "the probability density of X, given that $Y=y$."

If X and Y are independent, (26) simplifies to:

$$\mathbb{P}(X = x|Y = y) = f_X(x), \tag{27}$$

which corresponds to intuition: if X and Y are independent, the probability of "X taking on the value x" doesn't depend on what value Y has taken.

We record the following useful formula's:

$$\begin{aligned}\mathbb{P}(a < X < b \text{ and } c < Y < d) &= \int_a^b \int_c^d f_{X,Y}(x,y)dxdy \\ &= \int_a^b \int_c^d \mathbb{P}(X = x|Y = y) f_Y(y)dxdy \\ &= \int_c^d \left(\mathbb{P}(a \leq X \leq b|Y = y) f_Y(y)\right) dy\end{aligned}$$

Again, the concept of conditional probability density generalizes in a natural way from pairs of random variables to arbitrarily many random variables; only the notations get a bit more involved. For example, consider an N-tuple of rvs $X = (X_1, \cdots, X_N)$, with joint pdf $f_X = f_{X_1,\cdots,X_N}$, and pick a k, $1 \leq k \leq N$. If $x = (x_1, \cdots, x_N) \in \mathbb{R}^N$, we split x as

$$x = (x', x''),$$

with

$$x' = (x_1, \cdots, x_k) \in \mathbb{R}^k,$$

and

$$x'' = (x_{k+1}, \cdots, x_N) \in \mathbb{R}^{N-k}.$$

Similarly, we write

$$X = (X', X''),$$

where

$$X' = (X_1, \cdots, X_k),$$

and

$$X'' = (X_{k+1}, \cdots, X_N).$$

We can then write, symbolically, that $f_X = f_{X',X''}$. The *probability density of X', given that $X'' = x''$* is then equal to:

$$\mathbb{P}(X' = x'|X'' = x'') = \frac{f_X(x', x'')}{f_{X''}(x'')}, \tag{28}$$

where $f_{X''}(x'')$ is the marginal distribution of X'', obtained by integrating out the x'-variables:

$$\begin{aligned}f_{X''}(x'') &= \int_{\mathbb{R}^k} f_X(x', x'')dx' \\ &= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f_{X_1,\cdots,X_N}(x_1, \cdots, x_k, x_{k+1}, \cdots, x_N)dx_1 \cdots dx_k.\end{aligned}$$

Conditional densities are very useful in defining stochastic processes. For example, a so-called *Markov processes* can be specified by defining the conditional probability densities:

$$\mathbb{P}(X_t = x | X_s = y), \quad 0 \leq s < t,$$

together with the defining Markov property that, for any $s_1 < \cdots < s_N < t$,

$$\mathbb{P}(X_t = x | X_{s_1} = s_1, \cdots X_{s_N} = s_N) = \mathbb{P}(X_t = x | X_{s_N} = y_N).$$

The last equation is a mathematical way of stating that "the future only depends on the past via the present". The "transition probabilities" $\mathbb{P}(X_t = x | X_s = y)$ will have to satisfy certain consistency conditions, as we will see in the next chapter.

2.4. **The Central Limit Theorem.** In its simplest form, the central limit theorem is concerned with sums of rv which are *independent and identically distributed* (usually abbreviated as *iid*). The latter means all X_j have the same probability distribution functions: $F_{X_1} = F_{X_2} = \cdots$. We also require that their mean and variance are finite:

$$\mu := \int_{-\infty}^{\infty} x f_{X_j}(x) dx < \infty, \quad \sigma^2 := \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx < \infty;$$

note that since they are identically distributed, all X_j will identical mean and variance.

A short computation shows that the sums $S_N = X_1 + \cdots + X_N$ will have mean $N\mu$ and variance $N\sigma^2$, which both tend to infinity with N. We will therefore consider the *normalized sums:*

$$\hat{S}_N = \sum_{j=1}^{N} \frac{X_1 + \cdots + X_N - N\mu}{\sigma\sqrt{N}},$$

which have mean 0 and variance 1. The central limit theorem states that, for big N, these normalized sums will approximately be distributed according to the standard normal distribution, no matter what cdf of the X_j we started with, provided it was one having finite mean and variance. The precise formulation is as follows:

Theorem 2.28. *(Central Limit Theorem or CLT): Let $X_1, X_2, \cdots$ be a sequence of iid rvs having mean μ and variance σ^2. Then, for any real numbers $a < b$,*

$$\mathbb{P}\left(a < \hat{S}_N \leq b\right) \to \int_a^b e^{-x^2/2} \frac{dx}{\sqrt{2\pi}},$$

as $N \to \infty$. *$$*

We will not prove this theorem here. Historically, it was first proved in the case when the X_j were *iid Bernouilli*, meaning that X_j could take on two values, 1 and 0, say, with probabilities p and $1-p$. In this case, S_N has a binomial distribution, and one could use Stirling's formula for the asymptotic behavior of the binomial coefficients $\begin{pmatrix} N \\ k \end{pmatrix}$. This was a quite involved computation and a much slicker proof and more generally valid proof can be given using the concept of a characteristic function, which is also the basis of one of the modern proofs.

For the sums themselves, we infer that for big N:

$$\mathbb{P}\left(\sigma a\sqrt{N} + N\mu < S_N < \sigma b\sqrt{N} + N\mu\right) \simeq \int_a^b e^{-x^2/2}\frac{dx}{\sqrt{2\pi}},$$

or

$$\mathbb{P}\left(A < S_N < B\right) = \int_{(A-N\mu)/\sigma\sqrt{N}}^{(B-N\mu)/\sigma\sqrt{N}} e^{-x^2/2}\frac{dx}{\sqrt{2\pi}}.$$

The CLT explains why the normal distribution so often occurs in probabilistic modelling. When we want to model a random phenomenon which can be regarded as the sum of a lot of small independent but basically identically distributed random influences, the CLT suggests that it is reasonable to take a normally distributed rv for this. This is the basic intuition behind modelling financial asset returns by Gaussian rvs, as in the standard Geometric Brownian Motion model for stock-prices. One has to be careful, however: theorem 2.28 is for arbitrary but fixed a and b, and basically only tells us something about the behavior of the centre of the distribution of $\widehat{S}_N$. Indeed, empirical work during the 90's (and also much earlier) has shown that the actual stock returns have *fat-tailed* distributions, in the sense that very large or very small returns occur with much larger probabilities tan predicted by the normal model.

In the next chapter, we will use the CLT to introduce Brownian motion as a limit of a sequence of random walks, and from there go on to introduce the Ito calculus.

Exercise 2.29. *(moments of the normal distribution)* Let $Z \sim N(0,1)$ be a standard normal distribution, and let

$$m_n := \mathbb{E}(Z^n) = \int_{-\infty}^{\infty} x^n e^{-x^2/2}\frac{dx}{\sqrt{2\pi}},$$

be its n-th moment.

(a) Explain why all odd moments of Z are 0.

(b)Show that the even moments are related by $m_{2k} = (2k-1)m_{2k-2}$, and deduce from this that

$$m_{2k} = \frac{(2k)!}{2^k k!}.$$

(*int:* integrate by parts.)
(c) Now consider the odd moments of the *absolute value* $|Z|$ of Z:

$$\mathbb{E}(|Z|^{2k+1}) = \hat{m}_{2k+1}.$$

Show that as long as $2k-1 > 0$, $\hat{m}_{2k+1} = 2k\ \hat{m}_{2k-1}$, and deduce from this that:

$$\hat{m}_{2k+1} = 2^k k! \sqrt{\frac{2}{\pi}}.$$

$$

Exercise 2.30. Show that if $Z \sim N(0, \sigma^2)$, then $\sigma^{-1}Z \sim N(0,1)$. Use this and the previous exercise to compute the moments of Z and of $|Z|$. $$

3. **Brownian Motion**

3.1. **Introducing Brownian motion.** Brownian motion is a fundamental building block of modern quantitative finance. Indeed, the basic model for financial asset prices assumes that the log-returns follow a Brownian motion with drift. A convenient way to understand Brownian motion is by introducing it as a limit of random walks with ever smaller step sizes. We already encountered the random walk in the statement of the CLT: given an infinite sequence $(X_j)_{j\geq 0}$ if iid random variables, the sequence $(S_n)_{n=1,2,\cdots}$ of sums

$$S_n = X_1 + \cdots + X_n, \quad n = 1, 2, \cdots, \tag{29}$$

is called a *(general) random walk*. To fix ideas, people often consider the classical random walk in which X_j is a random variable which can only take the values ± 1, with equal probabilities:

$$X_j = \begin{cases} 1 & \text{with probability } 1/2 \\ -1 & \text{with probability } 1/2 \ . \end{cases} \tag{30}$$

Such rvs, with more generally $\mathbb{P}(X_j = 1) = p$ for an arbitrary $p \in [0, 1]$, are called Bernouilli random variables. The random walk can, somewhat whimsically, be thought of as the path of a drunkard, walking along a long road, who randomly takes either one step forward or one step backward at regularly spaced times $n = 1, 2, \cdots$. We will assume (30), for concreteness, although this is by no means necessary: having $\mathbb{E}(X_j) = 0$ and $\mathrm{Var}(X_j) = 1$ suffices for what follows.

We already know that

$$\mathbb{E}(S_n) = 0, \ \mathrm{VaR}\ (S_n) = n.$$

We will now consider a sequence of re-scaled random walks $(S_n^{(N)})_{n=1,2,\cdots}$, one for each $N \in \mathbb{N}$, having ever smaller step sizes $N^{-1/2}$ by replacing X_j by $X_j/\sqrt{N}$. We also imagine these steps to take place at times j/N instead of j. It is convenient to embed these random walks in a sequence of continuous time processes $(W_t^{(N)})_{t\geq 0}$, defined as follows: for a fixed N, divide the real half-line $\mathbb{R}_{\geq 0} = \{t : t \geq 0\}$ in intervals $[n/N, (n+1)/N)$ of size $1/N$. For any $t \geq 0$, we can find a unique positive integer n such that

$$\frac{n}{N} \leq t < \frac{n+1}{N},$$

and we put

$$W_t^{(N)} := \frac{X_1 + \cdots + X_n}{\sqrt{N}}. \tag{31}$$

This is fine if $t \geq 1/N$, since then $n \geq 1$. If $0 \leq t < 1/N$, we simply put $W_t^{(N)} := 0$, by definition.

One can think of $W_t^{(N)}$ as describing the return of an asset at time t, with trading taking place at equally spaced discrete times $1/N, 2/N, \cdots$, and price changes being limited so that returns can only jump by $\pm 1/\sqrt{N}$. (The reason for thinking of returns rather than prices themselves is because $W_t^{(N)}$ can be negative). Alternatively, a physicist might think of some small particle immersed in a 1-dimensional fluid, and undergoing a position change of $\pm 1/\sqrt{N}$ at times j/N, under the influence of molecular collisions.

We can write down (31) a little more compactly if we use the greatest integer part notation: if $x \geq 0$ is a positive real number, than its *greatest integer part* $[x]$ is, by definition, the largest positive integer which is $\leq x$. Thus, $[Nt] = n$ if $n/N \leq t < (n+1)/N$, and we can write

$$W_t^{(N)} = \frac{1}{\sqrt{N}} \sum_{j \leq [tN]} X_j, \tag{32}$$

putting the sum, by definition, equal to 0 if the set of j's over which we sum is empty (which is the case if $t < 1/N$). One easily checks that, if $t \in [n/N, (n+1)/N)$, then

$$\mathbb{E}(W_t^{(N)}) = 0, \quad \text{Var}\ (W_t^{(N)}) = \frac{n}{N} = \frac{[tN]}{N} \simeq t,$$

the latter approximation being valid for big N. Also, for a strictly positive t, if N is big, then $n = [tN]$ is big, so that, by the CLT, theorem 2.28, for any fixed t,

$$W_t^{(N)} = \sqrt{\frac{[tN]}{N}} \cdot \frac{1}{\sqrt{[tN]}} \sum_{j \leq [tN]} X_j, \tag{33}$$

will be approximately distributed as $t \cdot Z$, where $Z \sim N(0,1)$ is a standard normal rv. That is, $W_t^{(N)}$ will be approximately $N(0,t)$-distributed, for large N. At $t = 0$, we of course have $W_0^{(N)} = 0$, which is consistent with this. More generally, if $t > s$, let us consider the difference

$$W_t^{(N)} - W_s^{(N)} = N^{-1/2} \sum_{[sN]+1}^{[tN]} X_j. \tag{34}$$

One easily checks that this has variance $([tN] - [sN])/N \to t - s$ as $N \to \infty$, so that, for big N, the CLT implies that

$$W_t^{(N)} - W_s^{(N)} \text{ is approximately } N(0, t-s)\text{-distributed .} \tag{35}$$

Furthermore, if we take a third time, $u \leq s$, then all the X_j's which figure in the sum which defines $W_u^{(N)}$ have an index $j \leq [uN] < [sN] + 1$, and are therefore *independent* of the X_j's which occur in $W_t^{(N)} -$

$W_s^{(N)}$, simply since, by hypothesis, different X's are independent. We therefore have the property:

$$u \leq s < t \Rightarrow W_u^{(N)},\ W_t^{(N)} - W_s^{(N)} \text{independent.} \tag{36}$$

We now take the limit of $N \to \infty$. It seems reasonable to suppose that there exist a family of limiting rvs W_t, $t \geq 0$, such that $\mathbb{P}(W_t \leq x) = \lim_{N\to\infty} \mathbb{P}(W_t^{(N)} \leq x)$, or equivalently,

$$F_{W_t^{(N)}}(x) \to F_{W_t}(x), \quad N \to \infty,$$

having the following properties:

- (i) $W_0 = 0$.
- (ii) if $0 \leq s < t$, then $W_t - W_s \sim N(0, t-s)$.
- (iii) future changes are independent of past and present: if $0 \leq u \leq s < t$, then the rv's $W_t - W_s$ and W_u are independent.

One might also argue that, since the sizes of the jumps in $W_t^{(N)}$ tend to 0 as $N \to \infty$, the limiting process $(W_t)_{t\geq 0}$ should somehow be continuous in t: the idea is that for $\Delta t \sim 1/N$, $|\Delta W^{(N)}| \sim 1/\sqrt{N} \sim \sqrt{\Delta t} \to 0$ as $\Delta t \to 0$). We therefore expect that

- (iv) The paths $t \to W_t$ are continuous.

Definition 3.1. *(Brownian motion)* A stochastic process $(W_t)_{t\geq 0}$ which has properties (i), (ii) and (iii), and which also satisfies (iv), in the sense that, with probability 1, it has continuous sample paths, is called a Brownian motion. $$

The phrase 'having continuous sample paths with probability 1' has a precise technical sense which will be explained in detail in the second half of these lectures, after we will have had a look at measure theoretic probability.

There is a further property of Brownian motion which is plausible on the grounds of our construction, namely that Brownian motion paths $t \to W_t$, although continuous, are unlikely to be differentiable: if we consider $W_t^{(N)}$ on its natural scale of $\Delta t = 1/N$, then $\Delta W^{(N)}/\Delta t \simeq 1/\sqrt{\Delta t} \to \infty$ as $\Delta t \to 0$. This *non-differentiability* of Brownian motion paths can be shown to be a consequence of conditions (i) - (iv), and Brownian motion paths provide examples of continuous functions which are nowhere differentiable (do not have a tangent to their graphs at any point).

Some historical remarks. Brownian motion was introduced by L. Bachelier in his 1900 thesis on the mathematical modelling of the Parisian bourse. Although he did not quite proceed in the way presented here, in his model the $N^{-1/2}X_j$ would present daily price changes. He gave an interesting economical motivation for taking X_j with $\mathbb{E}(X_j) = 0$: he argued that, since typically somebody who wanted to sell a stock

could always find a buyer, and vice versa, there had to be about as many people expecting a given stock to go up as there would be expecting the stock's price to go down. Hence the (mathematical) expectation of the price change had to be 0. Also, his model assumed that future price changes only depend on the past through the present, which is formalized in condition (iii) above. Bachelier therefore already had a clear idea of what later on became known as the Efficient Market Hypothesis.

Unfortunately, Bachelier's work was somewhat sadly ignored at the time, and Brownian motion was independently rediscovered around 1905 by the physicists Einstein and Smoluchowski. In their work, Brownian motion is used as a model for the movement of microscopically small particles suspended in a (one-dimensional) fluid, under the influence of molecular collisions. Given that the particle is at position W_s at a time s, then, at a future time $t > s$, it will have changed position by an amount x with a probability of $(2\pi(t-s))^{-1/2}e^{-x^2/2(t-s)}$. It's mean position change is 0, but its standard deviation, which is a measure of it's deviation from this mean, grows with t like $\sqrt{t-s}$. This can be interpreted by saying that we are most likely to find the particle in an interval of size the order of $\sqrt{t-s}$, centered at where it was at time s, and this independently of what happened before s. Note that the distance covered grows with the square root of the time elapsed, instead of linearly with time. Such a motion is caled *diffusion*, whence the name of *diffusion processes* for the kind of stochastic processes used in Finance. In physics, a typical example of a diffusion phenomenon is the flow of heat. $$

*__Remark 3.2.__ There is a somewhat delicate mathematical point I glossed over above: how do we know that there actually *exist* a limiting family of random variables $(W_t)_{t\geq 0}$ with the stated properties (i), (ii) and (iii), let alone (iv)? Readers having a background in mathematics know that limits do not always exist under all circumstances. To answer this question we will need to be much more precise on what probabilities and random variables are, mathematically speaking. For Brownian motion, this was the step taken by N. Wiener in the 1920's, who gave the first mathematically rigorous construction of Brownian motion, and in whose honor this process is also often called the *Wiener process*, and denoted by W_t. We will examine these matters more closely in the second part of these lecturs. $$

3.2. **Simple properties of Brownian motion.** To get more of a feeling for Brownian motion, we compute some moments. These computations will turn out to be extremely useful in the next chapter, when we introduce Ito calculus. Let us fix a $t > 0$ and let $h > 0$ also. Then, by (ii), the mean of the increment $W_{t+h} - W_t$ clearly is 0:

$$\mathbb{E}(W_{t+h} - W_t) = 0. \tag{37}$$

The mean is 0 since $W_{t+h} - 0W_t$ has the same probability to be positive as to be negative. To get a better idea of the mean size of $W_{t+h} - W_t$, we compute the expectation of the *absolute value* of $W_{t+h} - W_t$:

$$\mathbb{E}(|W_{t+h} - W_t|) = (\text{constant})\sqrt{h}. \tag{38}$$

In fact, the left hand side is

$$\int_{-\infty}^{\infty} |x| e^{-x^2/2h} \frac{dx}{\sqrt{2\pi h}} = \sqrt{h} \int_{\infty}^{\infty} |z| e^{-z^2/2} \frac{dz}{\sqrt{2\pi}},$$

by the change of variables $y = \sqrt{h}z$, and the constant can be evaluated to be $\sqrt{2/\pi}$: see exercise 2.29,

We next look at the square of the increment, $(W_{t+h} - W_t)^2$. Using (ii) and the fact that the variance of a rv X equals $\text{Var}\,(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$, we find

$$\begin{aligned} \mathbb{E}\left((W_{t+h} - W_t)^2\right) &= \mathbb{E}\left((W_{t+h} - W_t)^2\right) - \left(\mathbb{E}(W_{t+h} - W_t)\right)^2 \\ &= \text{Var}\,(W_{t+h} - W_t) \\ &= h, \end{aligned} \tag{39}$$

where we used (37). Hence the the mean of $(W_{t+h} - W_t)^2$ is h. We also compute its variance, which is:

$$\begin{aligned} \text{Var}\left((W_{t+h} - W_t)^2\right) &= \mathbb{E}\left((W_{t+h} - W_t)^4\right) - h^2 \\ &= \int_{\mathbb{R}} x^4 e^{-x^2/2h} \frac{dx}{\sqrt{2\pi h}} - h^2 \\ &= 3^2 - 2h^2 = 2h^2, \end{aligned}$$

by an easy computation (or using exercises 2.29 and 2.30). Hence

$$\text{Var}\left((W_{t+h} - W_t)^2\right) = 2^2. \tag{40}$$

By formula (5), the expectation of any function $g(W_t)$ of Brownian motion can be evaluated as:

$$\mathbb{E}(g(W_t) = \int_{\mathbb{R}} g(x) e^{-x^2/2t} \frac{dx}{\sqrt{2\pi t}}. \tag{41}$$

As a further example we compute:

$$\begin{aligned} \mathbb{E}\left(e^{W_t}\right) &= \int_{-\infty}^{\infty} e^x e^{-x^2/2t} \frac{dx}{\sqrt{2\pi t}} \\ &= \int e^{-\frac{1}{2}(z^2 - 2\sqrt{t}z)} \frac{dz}{\sqrt{2\pi}} \text{ (changing variables } x = \sqrt{t}z \text{)} \\ &= e^{t/2} \int e^{-(z-\sqrt{t})^2/2} \frac{dz}{\sqrt{2\pi}} \text{ (completing the square)} \\ &= e^{t/2}. \end{aligned}$$

3.3. **Markov property and transition probabilities.** Brownian motion is a basic example of an important class of stochastic processes called *Markov processes*. To explain the notion of a Markov process, let us begin by looking at a general process, $(X_t)_{t\geq 0}$. To be able to compute interesting quantities associated to such a process, like for example the probability that for given times $t_1, \cdots, t_N$, X_{t_j} lies in given intervals $[a_j, b_j]$, we have to know all joint probability distributions of $(X_{t_1}, \cdots, X_{t_N})$, for arbitrary $t_1, \cdots, t_N \geq 0$. Here we can, without essential loss of generality, limit ourselves to t_j's such that $0 \leq t_1 < t_2 < \cdots < t_N$. We will assume that all these have probability densities, and we will systematically use the (mathematically disputable but conceptually very convenient) notation

$$\mathbb{P}(X_1 = x_1, \cdots, X_N = x_N)$$

for the joint pdf of the random vector $(X_1, \cdots, X_N)$ at a point $(x_1, \cdots, x_N)$. By the rules for conditional probabilities (cf. section 2.3) we have that, for example:

$$\mathbb{P}(X_{t_2} = x_2, X_{t_1} = x_1) = \mathbb{P}(X_{t_2} = x_2 | X_{t_1} = x_1)\mathbb{P}(X_{t_1} = x_1).$$

Similarly,

$$\begin{aligned} &\mathbb{P}(X_{t_3} = x_3, X_{t_2} = x_2, X_{t_1} = x_1) \\ &= \mathbb{P}(X_{t_3} = x_3 | X_{t_2} = x_2, X_{t_1} = x_1)\mathbb{P}(X_{t_2} = x_2, X_{t_1} = x_1) \\ &= \mathbb{P}(X_{t_3} = x_3 | X_{t_2} = x_2, X_{t_1} = x_1)\mathbb{P}(X_{t_2} = x_2 | X_{t_1} = x_1)\mathbb{P}(X_{t_1} = x_1), \end{aligned} \tag{42}$$

etc. Now suppose that our process is such that, for any $0 \leq t_1 < t_2 < \cdots < t_N$ we have that

$$\begin{aligned} &\mathbb{P}(X_{t_N} = x_N | X_{t_{N-1}} = x_{N-1}, \cdots, x_{t_1} = x_1) \\ &= \mathbb{P}(X_{t_N} = x_N | X_{t_{N-1}} = x_{N-1}). \end{aligned} \tag{43}$$

Intuitively, this means that the process has no memory: the probability that $X_{t_N} = x_N$ (or, more precisely, that it is in a small interval $[x_N, x_N + dx_N]$) depends only on the last recorded value, $X_{t_{N-1}} = x_{N-1}$, not on earlier ones. Assuming (43), formula (42) then simplifies to:

$$\begin{aligned} &\mathbb{P}(X_{t_3} = x_3, X_{t_2} = x_2, X_{t_1} = x_1) \\ &= \mathbb{P}(X_{t_3} = x_3 | X_{t_2} = x_2)\mathbb{P}(X_{t_2} = x_2 | X_{t_1} = x_1)\mathbb{P}(X_{t_1} = x_1), \end{aligned}$$

and, more generally[8] ,

$$\begin{aligned} &\mathbb{P}(X_{t_N} = x_N, \cdots, X_{t_1} = x_1) \\ &= \prod_{j=2}^{N} \mathbb{P}(X_{t_j} = x_j | X_{t_{j-1}} = x_{j-1}) \cdot \mathbb{P}(X_{t_1} = x_1). \end{aligned} \tag{44}$$

Definition 3.3. A stochastic process which satisfies (43) for arbitrary times $0 \leq t_1 < \cdots < t_N$ is called a *Markov process*. If $t < s$, the probability densities

$$p(x, t; y, s) := \mathbb{P}(X_s = y | X_t = x), \tag{45}$$

[8]The symbol $\prod$ stands for product: $\Pi_{j=1}^{k} a_j = a_1 a_2 \cdots a_k$.

are called the *transition probability densities*, or simply the *transition probabilities*. We will take $X_0 = x_0$, the position at time $t = 0$, as given.

$$

With this notation, (44) becomes:

$$\mathbb{P}(X_{t_N} = x_N, \cdots, X_{t_1} = x_1) = \prod_{j=1}^{N} p(x_{j-1}, t_{j-1}; x_j, t_j). \tag{46}$$

Remark 3.4. To remember the time-ordering of the variables in (45), it is helpful to read the left hand side as "$\mathbb{P}((x,t) \to (s,y))$", the transition probability of going from x at t to y at s, $s > t$. $$

Proposition 3.5. *Brownian motion $(W_t)_{t\geq 0}$ is a Markov process.*

Proof. Since $W_{t_N} - W_{t_{N-1}}$ is independent of $W_{t_1}, \cdots, W_{t_{N-1}}$, we have that

$$\begin{aligned}
& \mathbb{P}(W_{t_N} = x_N | W_{t_{N-1}} = x_{N-1}, \cdots, W_{t_1} = x_1) \\
= \; & \mathbb{P}(W_{t_N} - W_{t_{N-1}} = x_N - x_{N-1} | W_{t_{N-1}} = x_{N-1}, \cdots, W_{t_1} = x_1) \\
= \; & \mathbb{P}(W_{t_N} - W_{t_{N-1}} = x_N - x_{N-1}) \;\; \text{(by independence: cf. (27))} \\
= \; & \mathbb{P}(W_{t_N} - W_{t_{N-1}} = x_N - x_{N-1} | W_{t_{N-1}} = x_{N-1}) \;\; \text{(independence again)} \\
= \; & \mathbb{P}(W_{t_N} = x_N | W_{t_{N-1}} = x_{N-1}).
\end{aligned}$$

$$

Observe that this argument shows that the transition probabilities of Brownian motion are given by:

$$\begin{aligned}
\mathbb{P}(W_s = y | W_t = x) &= \mathbb{P}(W_s - W_t = y - x) \\
&= \frac{1}{\sqrt{2\pi(s-t)}} e^{-(y-x)^2/2(s-t)} \\
&=: p_0(y - x; s - t), \quad s > t \qquad (47)
\end{aligned}$$

where we have put

$$p_0(x,t) := \frac{1}{\sqrt{2\pi t}} e^{-x^2/2t};$$

also note that in fact $p_0(x,t) = \mathbb{P}(0,0|x,t)$, the transition probability for going from $(0,0) \to (x,t)$.

This gives us a first example of a very important connection between stochastics processes and partial differential equations or PDEs. Readers having a degree in physics or engineering will probably know that, for a fixed (y,s), $v(x,t) := p_0(y - x, s - t)$ as a function of (x,t), is a solution of the *backward heat equation:*

$$\frac{\partial v}{\partial t} + \frac{1}{2} \frac{\partial^2 v}{\partial x^2} = 0. \tag{48}$$

Similarly, $w(y,s) := p_0(y-x, s-t)$, as function of (y,s), satisfies the *heat equation:*

$$\frac{\partial w}{\partial s} - \frac{1}{2}\frac{\partial^2 w}{\partial y^2} = 0. \tag{49}$$

(If this is new for you, you should check it at least once in your life: see exercises.) The heat equation was introduced some 100 years before Brownian motion, by J. B. Fourier, who sought to describe the flow of heat in bodies. We will see in a later chapter that, under certain conditions, the transition densities of Markov processes satisfy certain PDEs which are called the backward and forward Kolmogorov equations, of which (48) and (49) are particular examples.

The great thing about Markov processes is that the transition probabilities fix all the joint probability densities of the process at arbitrary times, as shown by (44) , (46). For example, for Brownian motion we can now immediately write down that:

$$\begin{aligned} &\mathbb{P}(W_{t_N} = x_N, W_{t_{N-1}} = x_{N-1}, \cdots, W_{t_1} = x_1) \\ &p_0(x_N, t_N - t_{N-1}) p_0(x_{N-1} - x_{N-2}, t_{N-1} - t_{N-2}) \cdots p_0(x_1, t_1) \\ &= \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi(t_j - t_{j-1})}} e^{-(x_j - x_{j-1})^2/2(t_j - t_{j-1})}, \end{aligned} \tag{50}$$

where we' set $t_0 = x_0 = 0$ to make things agree for $j = 0$. Joint probabilities for being in consecutive intervals then follow by simple integration:

$$\mathbb{P}(a_1 < W_{t_1} < b_1, \cdots, a_N < W_{t_N} < b_N) =$$
$$\int_{a_1}^{b_1} \cdots \int_{a_N}^{b_N} \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi(t_j - t_{j-1})}} e^{-(x_j - x_{j-1})^2/2(t_j - t_{j-1})}\, dx_1 \cdots dx_N,$$

a formula which in some books is used to define Brownian motion directly.

3.4. **Brownian motion as a Gaussian process.** We first recall the definition of normal random vectors and multi-variate normal distributions. We will freely use matrix notation. Points $x = (x_1, \cdots, x_N)$ of $\mathbb{R}^N$ will be thought of as column vectors $(x_1\ x_2\ \cdots x_N)^t$, the 't' standing for transpose. Let $\mathbb{Z} = (Z_1, \cdots, Z_N)$ be a vector of independent normally distributed random variables Z_j with mean 0 and variance 1: $Z_j \sim N(0,1)$. If $\mu = (\mu_1, \cdots, \mu_N) \in \mathbb{R}$, and if $\mathbb{H} = (H_{ij})_{i,j}$ is an arbitrary $N \times N$-matrix, then the new random vector $\mathbb{X}$, defined by

$$\mathbb{X} = \mu + \mathbb{H}\mathbb{Z}, \tag{51}$$

is called a *normal random vector*. Its (component-wise computed) mean is clearly

$$\begin{aligned}\mathbb{E}(\mathbb{X}_i) &= \mu_i + \mathbb{E}\left(\sum_j H_{ij} Z_j\right) \\ &= \mu_i + \sum_j H_{ij}\mathbb{E}(Z_j) \\ &= \mu_i,\end{aligned}$$

since the expectation of Z_i is 0. As vectors,

$$\mathbb{E}(\mathbb{X}) = \mu.$$

The *variance-covariance matrix* of a random-vector $\mathbb{X}$ is defined as the symmetric matrix:

$$\mathbb{V} := \left(\mathbb{E}(X_i - \mu_i)(X_j - \mu_j) \right)_{1 \le i,j \le N}, \tag{52}$$

where $\mu_i = \mathbb{E}(X_i)$, the mean of X_i. If $\mathbb{X} = \mu + \mathbb{H}\,\mathbb{Z}$, then its variance-covariance matrix will be

$$\mathbb{V} = \mathbb{H}\mathbb{H}^t, \tag{53}$$

where the product on the right is the matrix product, and where the 't ' means again taking the transpose. Indeed,

$$\begin{aligned}\mathbb{E}\left((X_i - \mu_i)(X_j - \mu_j)\right) &= \mathbb{E}\left((\mathbb{H}\mathbb{Z})_i\,(\mathbb{H}\mathbb{Z})_j\right) \\ &= \sum_{k=1}^{N}\sum_{l=1}^{N} \mathbb{E}(H_{ik}H_{jl}Z_kZ_l) \\ &= \sum_{k=1}^{N} H_{ik}H_{jk} \\ &= \left(\mathbb{H}\mathbb{H}^t\right)_{ij},\end{aligned}$$

since, by independence,

$$\mathbb{E}(Z_kZ_l) = \delta_{kl} := \begin{cases} 1, \text{ if } k = l \\ 0, \text{ otherwise.}\end{cases}$$

If $\mathbb{H}$ is invertible (that is, if its matrix inverse $\mathbb{H}^{-1}$ exists, which is the case iff its determinant $\det(\mathbb{H}) \neq 0$), then one can compute that the joint pdf of $(X_1, \cdots, X_N)$ equals

$$f_{\mathbb{X}}(x) = \frac{1}{(2\pi)^{n/2}(\det\mathbb{V})^{1/2}} e^{-\left((x-\mu),\mathbb{V}^{-1}(x-\mu)\right)/2}, \tag{54}$$

where we've written $f_{\mathbb{X}}$ instead of $f_{X_1,\cdots,X_N}$, and where $\mathbb{V}$ was defined by (53). The proof uses the change of variables formula for multi-dimensional integrals and the Jacobian: cf. the appendix at the end of this chapter, after the exercises. Conversely, if $\mathbb{X}$ has the pdf (54)

(with $\mathbb{V}$ invertible), and if $\mathbb{H}$ is any matrix such that (53) holds, then one can show that

$$\mathbb{Z} := \mathbb{H}^{-1}(\mathbb{X} - \mu)$$

has pdf which is simply a product of standard normal distributions:

$$\mathbb{Z} \sim \frac{1}{(2\pi)^{n/2}} e^{-(z_1^2+\cdots+z_N^2)/2}.$$

The components of $\mathbb{Z}$ are therefore independent and identically $N(0,1)$-distributed, and it follows that $\mathbb{X} = \mu + \mathbb{H}\mathbb{Z}$ is Gaussian. Here we can always take $\mathbb{H} = \mathbb{V}^{1/2}$ (the matrix square root taken in spectral sense), or $\mathbb{H}$ given by the so-called *Cholesky decomposition* of $\mathbb{H}$, in which $\mathbb{H}$ can be taken to be either upper or lower triangular.

Observe that the pdf of a normal random vector is completely determined by its mean μ and its variance-covariance matrix $\mathbb{V}$. We say in this case that

$$\mathbb{X} \sim N(\mu, \mathbb{V}),$$

as an obvious clear generalization of the notation $X \sim N(\mu, \sigma^2)$ for single random variables.

It is pretty clear from (50) that, in case of a Brownian motion, each vector $(W_{t_1}, \cdots, W_{t_N})$ is normally distributed. Indeed, if we expand the squares in the exponent, we see that the joint pdf has the form

$$(2\pi)^{N/2} A \exp(-\sum a_{ij} x_i x_j),$$

for suitable $t_1, \cdots, t_N$-dependent numbers a_{ij} and A. The quadratic form in the exponent can be written as $(x, \mathbb{A}x)$, with $\mathbb{A} = (a_{ij})_{i,j}$, and the fact that the integral over all of $\mathbb{R}$ has to be equal to 1 forces A to be equal to the square root of the determinant of $\mathbb{A}$, as one sees by doing the change of variables $x = \mathbb{A}^{-1/2} y$ in the integral (alternatively, first do a rotation which brings $\mathbb{A}$ into diagonal form, and then re-scale each coordinate by the square root of the corresponding eigenvalue). This shows that $(W_{t_1}, \cdots, W_{t_N}) \sim N(0, \mathbb{A}^{-1})$. One should not try to read-off the variance-covariance matrix $\mathbb{V} = \mathbb{A}^{-1}$ from (50), since it is easier computed directly:

$$\mathbb{V}_{ij} = \mathbb{E}(W_{t_i}, W_{t_j}) = \min(t_i, t_j),$$

see the exercises.

Stochastic processes $(X_t)_{t\geq 0}$ having the property that *all finite dimensional pdfs* $f_{X_{t_1},\cdots,X_{t_N}}$ *are normally distributed with mean* 0 are called *Gaussian processes*. Such processes are completely determined by specifying the variance-covariance matrices

$$\big(\mathbb{E}(X_{t_i}\, X_{t_j})\big)_{1\leq i,j\leq N}.$$

Thus Brownian motion is a Gaussian process. Another important example of a Gaussian process is the so-called *Ornstein-Uehlenbeck process*, which we will encounter later on.

We end this section by observing an important property of normally distributed random vectors:

Proposition 3.6. *If $\mathbb{X} \sim N(0, \mathbb{V})$ with $\mathbb{V} = (V_{ij})_{i,j}$ diagonal: $V_{ij} = 0$ if $i \neq j$, then the components $X_1, \cdots, X_N$ of $\mathbb{X}$ are all independent.*

Proof. Indeed, the joint pdf becomes a simple product of the one-dimensional pdfs of the components:

$$F_{\mathbb{X}}(x) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi V_{jj}}} e^{-x_j^2/2V_{jj}}.$$

$ $

This proposition can be used to give yet another characterization of Brownian motion: see exercise 3.12.

3.5. **Exercises to chapter 3.**

Exercise 3.7. Let $(W_t)_{t\geq 0}$ be a Brownian motion, and fix two times $t < u$. Compute the following probability:

$$\mathbb{P}\left(W_t < 0, W_u > 0\right).$$

Simplify your answer as much as possible, and consider the special case of $u = 2t$.

Exercise 3.8. (a) Let $\xi \in \mathbb{R}$ be an auxiliary, non-random, variable taking values in $\mathbb{R}$. Show that

$$\mathbb{E}\left(e^{-\xi W_t}\right) = e^{t\xi^2/2}. \tag{55}$$

(b) The moments of $\mathbb{E}(W_t^n)$ can be computed using exercise 2.29. Give another derivation of these moments, by developing both sides of (55) in a power series in ξ, and comparing coefficients of like powers. $$

Exercise 3.9. Show that

$$\begin{aligned} \mathbb{E}\left(e^{\xi|W_t|}\right) &= 2e^{\xi^2 t^2/2}\left(1 - \Phi(-\xi\sqrt{t})\right) \\ &= 2e^{\xi^2 t^2/2}\Phi(\xi\sqrt{t}) \end{aligned}$$

where we recall that Φ is the cumulative normal distribution function:

$$\Phi(x) = \int_{-\infty}^{x} e^{-y^2/2}\frac{dy}{\sqrt{2\pi}}.$$

$$

Exercise 3.10. Show that if $(W_t)_{t\geq 0}$ is a Brownian motion, and if $a > 0$ is a positive real number, then the new process $(\widetilde{W}_t)_{t\geq 0}$ defined by:

$$\widetilde{W}_t = \frac{1}{\sqrt{a}} W_{at},$$

is also a Brownian motion. This is called the *self-similarity property* of Brownian motion. $ $

Exercise 3.11. Show that if $(W_t)_{t\geq 0}$ is a Brownian motion, then

$$\mathrm{Cov}(W_t, W_s) = \min(s, t).$$

(*Hint:* if, for example, $s < t$, write $W_t = W_s + (W_t - W_s)$ and use property (iii) of a Brownian motion). $$

Exercise 3.12. Conversely, let $(B_t)_{t\geq 0}$ be a mean 0 Gaussian stochastic process (with continuous sample paths), such that $B_0 = 0$, and such that $\mathrm{Cov}(B_t, B_s) = \min(s, t)$, for all $t, s \geq 0$. Show that $(B_t)_{t\geq 0}$ is a Brownian motion. $$

Exercise 3.13. As an application of the previous exercise, show that if $(W_t)_{t\geq 0}$ is a Brownian motion, then the new process $(\widehat{W})_{t\geq 0}$, defined for $t > 0$ by

$$\widehat{W}_t := tW_{1/t},$$

while $\widehat{W}_0 := 0$, is also a Brownian motion. $$

Exercise 3.14. Define the function $p_0 = p_0(x, t)$ by

$$p_0(x, t) = \frac{1}{\sqrt{2\pi t}} e^{-x^2/2t}, \quad x \in \mathbb{R}, t > 0.$$

Show that p_0 satisfies the *heat equation* on $t > 0$, that is:

$$\frac{\partial p_0}{\partial t} = \frac{1}{2} \frac{\partial^2 p_0}{\partial x^2}$$

Deduce from this that, for fixed s and y, $v(x, t) = p_0(y-x, s-t)$ (which is exactly the transition probability density for Brownian motion) satisfies the backward heat equation (48) on $t < s$. Also show that, as a function of y and s, $w(s, y) = p_0(y-x, s-t)$ satisfies the heat equation:

$$\frac{\partial w}{\partial s} - \frac{1}{2} \frac{\partial^2 w}{\partial y^2} = 0.$$

$ $

3.6. * **Appendix to Chapter 2: Proof of** (54)**.** We first quickly review the *change of variables formula for multiple integrals.* Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function, $H : \mathbb{R}^n \to \mathbb{R}^n$ a one-to-one map of $\mathbb{R}^n$ into itself, and let $F \subset \mathbb{R}^n$ be some a domain in $\mathbb{R}^n$ over which we want to integrate, for example an n-dimensional rectangle $[a_1, b_1] \times \cdots \times [a_n, b_n]$. Suppose that $H(x)$ has components:

$$H(x) = (H_1(x), \cdots, H_n(x)),\ x = (x_1, \cdots, x_n).$$

Then one defines the *Jacobian* of H by:

(56)

$$J_H(x) = \left| \det \begin{pmatrix} \partial H_1(x)/\partial x_1 & \partial H_1(x)/\partial x_2 & \cdots & \partial H_1(x)/\partial x_n \\ \partial H_2(x)/\partial x_1 & \partial H_2(x)/\partial x_2 & \cdots & \partial H_2(x)/\partial x_n \\ \vdots\vdots & \ddots & \vdots & \\ \partial H_n(x)/\partial x_1 & \partial H_n(x)/\partial x_2 & \cdots & \partial H_n(x)/\partial x_n \end{pmatrix} \right| ,$$

and we have the following change of variables formula:

$$\int_{H(F)} f(x)dx = \int_F f\left(H(y)\right)\ J_H(y)\ dy. \tag{57}$$

If $n = 1$, and $F = [a, b]$, this reduces to the well-known substitution formula:

$$\int_{H(a)}^{H(b)} f(x)dx = \int_a^b f(H(y))\ |H'(y)|\ dy,$$

obtained by substituting $x = (y)$; this formula holds provided that $H'(y) \neq 0$ for $a \leq y \leq b$.

An important example is when H is a linear map of $\mathbb{R}^n$, defined by a matrix $\mathbb{H}$:

$$H(x) := \mathbb{H}x := \left(\sum_{j=1}^n H_{1j}x_j, \cdots, \sum_{j=1}^n H_{nj}x_j \right).$$

In this case one easily sees that the matrix in (56) is simply $\mathbb{H}$, and that therefore:

$$J_H = |\det(\mathbb{H})|.$$

As an application we show how to derive the pdf of normal random vector:

Theorem 3.15. *Let $\mathbb{Z}$ be an $N(0, I)$-random vector. If $\mu \in \mathbb{R}^n$ and if $\mathbb{H}$ is an invertible matrix, then the new random vector*

$$\mathbb{X} = \mu + \mathbb{H}\mathbb{Z}, \tag{58}$$

is $N(\mu, \mathbb{V})$-distributed, where

$$\mathbb{V} = \mathbb{H}\mathbb{H}^t.$$

Conversely, each random vector $\mathbb{X} \sim N(\mu, \mathbb{V})$ can be written as (58), with $\mathbb{H}$ satisfying $\mathbb{H}\mathbb{H}^t = \mathbb{V}$, and $\mathbb{Z} \sim N(0, I)$.

Proof. Let $F = [a_1, b_1] \times \cdots \times [a_n, b_n]$ be an n-dimensional rectangle (we could, using measure-theoretic terminology, allow F to be any Borel subset of $\mathbb{R}^n$). Then, by the n-dimensional change of variables formula, putting $\mu + \mathbb{H}z = x$,

$$\begin{aligned}
\mathbb{P}(\mathbb{X} \in F) &= \mathbb{P}(\mu + \mathbb{H}z \in F) \\
&= \int_{\{z:\mu+\mathbb{H}z\in F\}} e^{-(z,z)/2} \frac{dz}{(2\pi)^{n/2}} \\
&= \int_F e^{\left(\mathbb{H}^{-1}(x-\mu),\mathbb{H}^{-1}(x-\mu)\right)/2} \frac{dx}{(2\pi)^{n/2}\det(\mathbb{H})} \\
&= \int_F e^{-\left(x-\mu,\mathbb{V}^{-1}(x-\mu)\right)/2} \frac{dx}{(2\pi)^{n/2}(\det(\mathbb{V}))^{1/2}},
\end{aligned}$$

since

$$\mathbb{V} = \mathbb{H}\mathbb{H}^t \Rightarrow \mathbb{V}^{-1} = \left(\mathbb{H}^{-1}\right)^t \mathbb{H}^{-1},$$

and $\det(\mathbb{V}) = \det(\mathbb{H})^2$.

Conversely, if $X \sim N(\mu\mathbb{V})$, then similar computations show that if $\mathbb{V} = \mathbb{H}\mathbb{H}^t$, then

$$\mathbb{Z} := \mathbb{H}^{-1}(\mathbb{X} - \mu) \text{ is } \ N(0, I),$$

and clearly $\mathbb{X} = \mu + \mathbb{H}\mathbb{Z}$. QED

4. A Crash Course in Ito Calculus

In this chapter we introduce the basic rules of stochastic calculus, also known as Ito calculus. We will do this using an intuitive approach which is based on calculus-style differentials, postponing the mathematically rigorous approach (founded on the Ito stochastic integral), to a later chapter.

4.1. **Stochastic differentials.** Brownian motion is not differentiable in the usual sense. As an indication of this, observe that, by (38),

$$\lim_{h\to 0} \mathbb{E}\left(\frac{|W_{t+h} - W_t|}{h}\right) = \lim_{h\to 0} \frac{\text{const.}}{\sqrt{h}} = \infty,$$

which makes it unlikely that the derivative with respect to time, defined as is usual by

$$\frac{d}{dt}W_t = \lim_{h\to 0} \frac{W_{t+h} - W_t}{h},$$

should exist. In fact, it can be shown that Brownian sample paths are nowhere differentiable with probability 1. Still, we would like to be able to talk about infinitesimally small increments of Brownian motion over an infinitesimal time-interval, from $[t, t + dt]$:

$$dW_t = W_{t+dt} - W_t. \tag{59}$$

Here dt is a calculus-style infinitesimal:

$$dt \neq 0, (dt)^2 = (dt)^3 = \cdots = 0,$$

usually interpreted as a number so small that higher powers can be neglected. There will be some important differences with ordinary calculus, though: first of all, dt will have to be a *positive* differential: $dt > 0$, so that (59) is a *differential into the future*, which is unknown at time t and on which we have, at best, probabilistic information. Furthermore, we will also encounter fractional powers of dt: $dt^{1/2}, dt, dt^{3/2}$, etc. However, as in calculus, powers higher than 1 are still to be neglected: $dt^{3/2} = dt^{5/2} = 0$, etc.

The key to the correct interpretation of dW_t is to remember condition (ii) of definition 3.1 of a Brownian motion, and to interpret $dW_t = W_{t+dt} - W_t$ as a normal rv with mean 0 and (infinitesimally small) variance dt:

$$dW_t \sim N(0, dt) \quad (dt > 0). \tag{60}$$

To get an idea of the size of dW_t, we apply (38) with $h = dt$:

$$\mathbb{E}(|dW_t|) = \text{const.}\sqrt{dt}.$$

This tells us that dW_t is a differential of size $\sqrt{dt}$ which, for very small dt, though small, is very much bigger that dt itself: $\sqrt{dt} >> dt$.

Next, we look at $(dW_t)^2$, which will be of size comparable to dt. Indeed, by (39),

$$\mathbb{E}\left((dW_t)^2\right) = dt.$$

Furthermore, by (40), the variance of this rv is:

$$\mathrm{Var}((dW_t)^2) = (dt)^2. \tag{61}$$

Now this is interesting: since we decided to neglect higher powers of dt, (61) tells us that $(dW_t)^2$ is a rv with variance 0, so not at all a rv anymore, but an ordinary (non-stochastic or deterministic) number whose value necessarily has to be equal to its mean:

$$(dW_t)^2 = dt \tag{62}$$

What about $dtdW_t$? By the above computations, it's mean is:

$$\mathbb{E}(dtdW_t) = dt\mathbb{E}(dW_t) = 0,$$

and it's variance is 0 also:

$$\mathbb{E}(dt^2 dW_t^2) = dt^3 = 0.$$

Hence, for the same reason as before, we have that

$$dt\ dW_t = 0, \quad dt^2 = 0. \tag{63}$$

Finally, we already know that $(dt)^2 = 0$, and as an immediate consequence of these relations, other combinations of powers of dW_t and dt will also count as 0: for example,

$$(dt)^3 dW_t = 0,\ (dt)^2 dW_t = 0,$$

and

$$dt\ (dW_t)^2 = (dt)^2 = 0,\ (dW_t)^3 = dW_t\ dt = 0,$$

etc.

Formula's (62) and (63) are the basic rules for *Ito's stochastic differential calculus*, and they are often summarized in the form of *Ito's multiplication table*:

$\cdot$	dW_t	dt
dW_t	dt	0
dt	0	0

(64)

Example 4.1. To see how this works in practice, let us compute the differential of the square of a Brownian motion, $d(W_t^2)$:

$$\begin{aligned} d(W_t^2) &= (W_{t+dt})^2 - W_t^2 \\ &= (W_t + dW_t)^2 - W_t^2 \\ &= \left(W_t^2 + 2W_t dW_t + (dW_t)^2\right) - W_t^2 \\ &= 2W_t dW_t + (dW_t)^2 \\ &= 2W_t dW_t + dt, \end{aligned}$$

where in the last line we used Ito's rules (64). $ $

4.2. **Review of the Taylor expansion.** We will want to generalize the preceding example to more general functions of W_t instead of just the square. For this we need to review a few basic facts concerning *Taylor expansions* of sufficiently differentiable functions. The simplest case is that of a function of a single variable, $f = f(x)$. Suppose f is $(k+1)$-times continuously differentiable, meaning that f has $(k+1)$ derivatives, which are all continuous functions[9]. As usual, we denote the successive derivatives of f by

$$f'(x) = f^{(1)}(x) = \frac{d}{dx}f(x),\ f''(x) = f^{(2)}(x) = \frac{d^2}{dx^2}f(x),$$

and in general

$$f^{(j)}(x) = \frac{d^j}{dx^j}f(x) := \frac{d}{dx}f^{(j-1)}(x).$$

The Taylor expansion of f around a point x approximates the value of f in a nearby point $x+h$ by a polynomial in h whose coefficients depend in a simple way on f's derivatives in x:

$$f(x+h) = f(x) + f'(x)h + \frac{1}{2!}f''(x)h^2 + \cdots + \frac{f^{(k)}(x)}{k!}h^k + O(|h|^{k+1}),$$

or, written more concisely,

$$f(x+h) = \sum_{j=0}^{k} \frac{f^{(j)}(x)}{j!}h^j \ + \ O(|h|^{k+1})$$

with $f^{(0)} := f$. Here the symbol $O(h^{k+1})$ means that the error which we make can be bounded by a constant times h^{k+1}, and is thus of smaller order than all previous terms in the expansion of $f(x+h)$, if h is small: $|h| \ << \ 1$.

We will in fact only need the Taylor expansion up to order $k = 2$, which reads:

$$f(x+h) = f(x) + f'(x)h + \frac{1}{2!}f''(x)h^2 + 0(h^3). \tag{65}$$

The Taylor formula generalizes to functions of several variable: $f = f(x_1, \cdots, x_n)$, defined on $\mathbb{R}^n$. We only state the case of the second order expansion, which again is all which will be needed here, and refer to any good multi-variable calculus book for the general case. If $h = (h_1, \cdots, h_n)$, then

$$\begin{aligned} f(x+h) \ &= \ f(x) \ + \ \sum_j \frac{\partial f}{\partial x_j}(x) \ + \\ &\qquad \frac{1}{2}\sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(x)h_i h_j + O(|h|^3), \end{aligned} \tag{66}$$

[9]This is not the minimal condition what follows, but suffices for all applications we'll consider.

where $|h| = (h_1^2 + \cdots + h_n^2)^{1/2}$, the Euclidean norm of h. The derivatives here are the usual partial derivatives of functions of more than one variable: for example, $\partial f/\partial x_1$ means that one takes the derivative w.r.t. x_1, considering $x_2, \cdots, x_n$ as constants, while $\partial^2 f/\partial x_1 \partial x_2$ means taking the derivative w.r.t. x_2 of the function $\partial f/\partial x_1(x_1, \cdots, x_n)$, regarding $x_1, x_3, \cdots, x_n)$ as constants, etc. Under rather mild conditions[10] on f, mixed derivatives can be computed in arbitrary order: $\partial^2 f(x)/\partial x_1 \partial x_2 = \partial^2 f(x)/\partial x_2 \partial x_1$, etc.

4.3. **Ito's lemma.** Let us now consider an arbitrary function of Brownian motion:

$$f(W_t),$$

$f = f(w)$ being a 3 times continuously differentiable function of $w \in \mathbb{R}$. We would like to know by how much $f(W_t)$ will change along a given Brownian motion path, when going one infinitesimal time-step into the future, from t to $t + dt$. That is, we'd like to compute

$$df = df(W_t) = f(W_{t+dt}) - f(W_t\ , t);$$

note that this will in general be a stochastic quantity. The idea is to simply Taylor expand f around (W_t): since $W_{t+dt} = W_t + dW_t$, we have that, by (65) with $x = W_t$ and $h = dW_t$

$$\begin{aligned} f(W_{t+dt}) &= f(W_t + dW_t) \\ &= f(W_t) + f'(W_t)dW_t + \frac{1}{2}f''(W_t)(dW_t)^2 + O\left((dW_t)^3\right) \\ &= f(W_t) + f'(W_t)dW_t + \frac{1}{2}f''(W_t)dt, \end{aligned}$$

by Ito's rules (64). Subtracting $f(W_t)$, we find:

Ito's lemma: simplest case *If $f = f(w)$ is thrice continuously differentiable, then*

$$df(W_t) = f'(W_t)dW_t + \frac{1}{2}f''(W_t)dt. \tag{67}$$

Example 4.2. *(Example 4.1 revisited)* Take $f(w) = w^2$. Then $f'(w) = 2w$, and $f''(w) = 2$, so that, by (67),

$$d(W_t^2) = 2W_t dW_t + dt,$$

in agreement with what we found before. $ $

Observe that (67) not only contains a dW_t, but als a dt, and to have a smoothly running and easily applicable formalism, we have to go a bit beyond (67), and consider functions of both Brownian motion and time:

$$f(W_t, t),$$

[10] e.g. continuous partial derivatives of the relevant order

where $f = f(w,t)$ is an ordinary (non-stochastic) function of 2 variables, w (for which we will substitute Brownian motion) and time t. We follow the same strategy as before, but we now use the 2-variable Taylor expansion, (66) with $n = 2$ and $(x_1, x_2) = (w,t)$. Using the shorthand notation

$$\partial_w = \frac{\partial}{\partial w}, \quad \partial_t = \frac{\partial}{\partial t},$$

for the partial derivatives, we easily find that

$$\begin{aligned} f(W_{t+dt}\ , t+dt) &= f(W_t\ , t) + \partial_w f(W_t\ , t)dW_t + \partial_t f(W_t\ , t)dt + \\ &+ \frac{1}{2}(\partial_w^2 f(W_t\ , t)dW_t^2 + 2\partial_{tw}^2 f(W_t\ , t)dtdW_t \\ &+ \partial_t^2 f(W_t\ , t)dt^2) + O\left(\left((dW_t)^2 + (dt)^2\right)^{3/2}\right). \end{aligned}$$

Now, Ito's multiplication rules (64) imply that $dtdW_t = (dt)^2 = 0$, while $((dW_t)^2 + (dt)^2)^{3/2} = (dt + (dt)^2)^{3/2} << (dt)^{3/2} = 0$, and we are simply left with:

$$df = \left(\partial_t f + \frac{1}{2}\partial_w^2 f\right) dt + (\partial_w f)dW_t,$$

both sides to be evaluated in the same point (W_t, t). This is the famous *Ito lemma*:

Theorem 4.3. *(Ito's lemma: functions of Brownian motion and time) Let $f = f(w,t) : \mathbb{R} \to \mathbb{R}$ be a 3 times continuous differentiable function. Then the infinitesimal change of $f(W_t, t)$ along a Brownian motion path is given by:*

$$df(W_t, t) = \left(\frac{\partial f}{\partial t} + \frac{1}{2}\frac{\partial^2 f}{\partial w^2}\right) dt + \frac{\partial f}{\partial w} dW_t, \tag{68}$$

all derivatives of f to be evaluated in the point (W_t, t). $$

Remarks 4.4. (i) rather than just memorizing (68), it pays just to remember how to derive it from the Ito rules (64).

(ii)* Formula (68) remains true under the weaker condition that f has continuous partial derivatives of order 1 and 2. One can even allow functions having singularities in their second, and even first derivatives, provided one interprets all terms in the correct way. This is connected with the concept of Tanaka's local time, which we won't go into; see the more advanced literature, like the book by Revuz and Yor. $ $

Examples 4.5. (i) Applying (68) (or even (67)) to $f = e^{\sigma w}$, we find:

$$d(e^{\sigma W_t}) = \frac{1}{2}\sigma^2 e^{\sigma W_t} dt + \sigma e^{\sigma W_t} dW_t. \tag{69}$$

Again note that this is different from the answer ordinary calculus leads you to expect: $d(e^{\sigma w}) = \sigma e^{\sigma w} dw$ if w is an ordinary (non-stochastic)

variable. Additional factors involving $\sigma^2/2$ are ubiquitous in Finance, and usually point to Ito's lemma having been applied somewhere.

(ii) As a slight variation on the previous example, consider

$$X_t = e^{-\frac{\sigma^2}{2}t+\sigma W_t}. \tag{70}$$

An easy computation using (68) now shows that:

$$dX_t = \sigma X_t dW_t. \tag{71}$$

This is our first example of a *stochastic differential equation* or *SDE*, and as such the stochastic analogue of the ordinary differential equation (or *ODE*),

$$dx_t = \sigma x_t dt, \tag{72}$$

which characterizes the exponential function. Indeed, the solution of (72) with initial value $x_0 = 1$ is the exponential function: $e^{\sigma t}$. In analogy with this, (70) is sometimes called (especially in the more mathematical literature), the *Doléans-Dade exponential* of W_t. $$

4.4. **Ito processes.** For any stochastic process $(X_t)_t$ one may consider its infinitesimal change into the future:

$$dX_t := X_{t+dt} - X_t, \quad dt > 0.$$

Definition 4.6. *(Ito process, informal definition)* $(X_t)_{t\geq 0}$ is called an *Ito process* if dX_t is related to an underlying Brownian motion $(W_t)_{t\geq 0}$ by:

$$dX_t = a_t dt + b_t dW_t, \tag{73}$$

where $(a_t)_{t\geq 0}$ and $(b_t)_{t\geq 0}$ are two auxiliary stochastic processes having the important property that *a_t and b_t only depend on the Brownian motion through its past values W_s, $s \leq t$*:

$$a_t,\ b_t = \text{Functions of } \left((W_s)_{s\leq t}, t\right). \tag{74}$$

$ $

*__Remark 4.7.__ For reasons which will become clear later on, to guarantee existence of such a process X_t it is extremely important that in particular b_t should not depend on any *future values* $W_{t'}$, $t' > t$ of W_t. To be more precise, we also need some kind of boundedness condition of a_t, b_t: simply requiring them to be uniformly bounded,

$$|a_t|, |b_t| \leq C < \infty,$$

with $C > 0$ independent of t for the range of times we're interested in, will certainly do. $ $

Example 4.8. Brownian motion itself is an Ito process: simply take $a_t = 0$ and $b_t = 1$. More generally, any $X_t = f(W_t, t)$ is an Ito process by Ito's lemma 4.3. Indeed, by (68),

$$dX_t = df(W_t, t) = a_t dt + b_t dW_t,$$

where

$$a_t = \frac{\partial f}{\partial t}(W_t, t) + \frac{1}{2}\frac{\partial^2 f}{\partial w^2}(W_t, t),$$

and

$$b_t = \frac{\partial f}{\partial w}(W_t, t).$$

Observe that these satisfy condition (74), being functions of t and W_t. $ $

A natural question now is whether functions of Ito-processes are again Ito-processes. The following extension of lemma 4.3 show that the answer is 'yes'. Its proof doesn't need any new ideas, but follows the same pattern as before.

Theorem 4.9. *(Ito's lemma for functions of Ito processes) Let $f = f(x,t)$ be thrice continuously differentiable. Then:*

$$\begin{aligned} df(X_t,t) &= \frac{\partial f}{\partial t}dt + \frac{\partial f}{\partial x}dX_t + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}(dX_t)^2, \\ &= \left(\frac{\partial f}{\partial t} + a_t\frac{\partial f}{\partial x} + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}b_t^2\right)dt + \frac{\partial f}{\partial x}b_t dW_t, \end{aligned} \tag{75}$$

where all derivatives on the right are to be evaluated in (X_t, t).

Remark 4.10. This is the form of Ito's lemma which is used most in applications.

Proof. Ito's multiplication table(64), and $dX_t = a_t dt + b_t dW_t$ easily imply that $dX_t dt = (dt)^2 = 0$, and the first line of (75) follows, as before, from a second order Taylor expansion of $f(x,t)$ around (X_t, t). The second line follows from the first simply by substituting $dX_t = a_t dt + b_t dW_t$, and by observing that

$$(dX_t)^2 = (a_t dt + b_t dW_t)(a_t dt + b_t dW_t) = b_t^2 dt,$$

again by Ito's multiplication table. $ $

4.5. **Stochastic differential equations.** Like in ordinary calculus, having differentials, one can consider equations between differentials, or *differential equations*. Equations involving the stochastic differentials dW_t are, unsurprisingly, known as *stochastic differential equations* or SDE's. These occur all the time in continuous time finance, and we need to have some idea how to solve them, both analytically (where possible) and numerically.

A typical (first order) SDE looks like:

$$dX_t = a(X_t, t)dt + b(X_t, t)dW_t \tag{76}$$

where $a = a(x,t)$ and $b = b(x,t)$ are given functions, and we are looking for a solution X_t of (76) subject to some, prescribed, *initial condition:*

$$X_{t=0} = X_0 \text{ given;} \tag{77}$$

here X_0 can be an ordinary real number, but also a rv.

In what sense do we wish to solve (76)? We would like to have a solution X_t (which in itself is going to be a rv) which can be expressed in terms of the initial condition X_0 and of the given Brownian motion $(W_s)_{s\geq 0}$ *up till time* t: that is, we are looking for solutions

$$X_t = \text{Function}\,(X_0, W_s(s \leq t))\,. \tag{78}$$

It is important that the solution at time t is only allowed to depend on past values $(W_s)_{s\leq t}$ of Brownian motion. In a financial context, where the W_t might represent something like a return, this makes sense: only past returns will have been observed: the future ones are as yet unknown, and we are only interested in solutions which can be computed on the basis of our present knowledge. We look at two important examples of analytically solvable SDEs.

Example 4.11. *(Geometric Brownian motion as a model for stock prices)* A simple model for stock-prices S_t is obtained by assuming that the return $(S_{t+dt} - S_t)/S_t$ over an infinitesimal period $[t, t+dt]$ is normally distributed, with mean and variance both proportional to the time interval dt:

$$\frac{dS_t}{S_t} \sim N(\mu dt, \sigma^2 dt), \quad \mu,\, \sigma^2 \text{ constant}\,.$$

Now, since $dW_t \sim N(0, dt)$, it follows that $\mu dt + \sigma dW_t \sim N(\mu dt, \sigma^2 dt)$ also[11] We therefore take as our model:

$$\frac{dS_t}{S_t} = \mu dt + \sigma dW_t, \tag{79}$$

or

$$dS_t = \mu S_t dt + \sigma S_t dW_t. \tag{80}$$

This is a SDE for the stock price S_t. Given an initial value S_0, it's solution is given by

$$S_t = S_0 e^{(\mu - \sigma^2/2)t + \sigma W_t}\,: \tag{81}$$

this can be checked by applying Ito's lemma (68) to compute the differential of the right hand side. The occurrence of the $(-\sigma^2/2)t$-term in the exponential on the right is explained by the $\partial_x^2 f/2$-term in (68): see also (69) above. The process S_t is called *geometric Brownian motion with drift* μ *and volatility* σ^2.

[11] We are using here that if $Z \sim N(0, a^2)$, then $\mu + \sigma Z \sim N(\mu, \sigma^2\, a^2)$.

Another way of computing the solution to (80) is by first observing that Ito's lemma 4.9 implies that

$$\begin{aligned} d\log S_t &= \frac{dS_t}{S_t} - \frac{1}{2}\frac{(dS_t)^2}{S_t^2} \\ &= (\mu - \frac{\sigma^2}{2})dt + \sigma dW_t, \end{aligned}$$

where we used (80) for the last line. This integrates to:

$$\log S_t = \log S_0 + (\mu - \frac{\sigma^2}{2})t + \sigma W_t,$$

which gives (81) upon taking the exponential of both sides. $$

Example 4.12. *(the Ornstein-Uehlenbeck or mean reverting process)* The Ornstein-Uehlenbeck SDE is given by

$$dX_t = \alpha(\theta - X_t)dt + \sigma dW_t. \tag{82}$$

To solve this, put $Y_t = X_t - \theta$. Then $dY_t = dX_t$ and so:

$$dY_t = -\alpha Y_t + \sigma dW_t$$

To get rid of the $-\alpha Y_t$, multiply Y_t by $e^{\alpha t}$; then

$$\begin{aligned} d(e^{\alpha t}Y_t) &= e^{\alpha t}dY_t + \alpha e^{\alpha t}Y_t \\ &= \sigma e^{\alpha t}dW_t, \end{aligned} \tag{83}$$

by the previous equation. Formally integrating from 0 to t, we would find that

$$e^{\alpha t}Y_t = Y_0 + \int_0^t \sigma e^{\alpha s}dW_s$$

or, remembering what Y_t stands for,

$$X_t = \theta(1 - e^{-\alpha t}) + X_0 e^{-\alpha t} + \sigma \int_0^t e^{\alpha(s-t)}dW_s. \tag{84}$$

Note, that here the solution up till time t depends on all W_s for times $s \leq t$. $$

Of course, (84) begs the question of what we precisely mean by the integral which occurs in the left hand side, and more generally by an integral of the type

$$\int_0^t f(s)dW_s. \tag{85}$$

That is, *what does it mean to integrate w.r.t. Brownian motion*? This is a point which we will have to address, before discussing the properties of the solution (84).

4.6. **Stochastic Integrals.** How can we give a sense to (85)? Let us again try to take our inspiration from ordinary calculus, which after all is a model for what we are trying to do. If $f = f(t)$ and $g = g(t)$ are deterministic, non-random, functions on the real line, a first natural idea is to interpret

$$\int_0^t f(s)dg(s), \tag{86}$$

as being

$$\int_0^t f(s)g'(s)ds, \tag{87}$$

where $g'(s) = dg(s)/ds$ is the derivative of g. Can this work for (85), that is, can we write:

$$\int_0^t f(s)\frac{dW_s}{ds}ds?$$

For this, one would have to be able to differentiate Brownian motion W_t with respect to time t, and we already argued that this is not possible. However, there is a way out if we recall the important, and extremely useful, trick of *integration by parts*: assuming that f and g are both continuously differentiable, we have that:

$$\begin{aligned}\int_0^t f(s)dg(s) &= \int_0^t f(s)g'(s)ds \\ &= [f(s)g(s)]_0^t - \int_0^t g(s)f'(s)ds \\ &= (f(t)g(t) - f(0)g(0)) - \int_0^y g(s)df(s).\end{aligned} \tag{88}$$

Now suppose that f is (continuously) differentiable, but g is not. Then we can still give a sense to the left hand side of (86), by *defining* it to be the right hand side of (88)! This works out all right, since the right hand side is perfectly well-defined under these conditions on f and g. In particular, we can now apply this definition to (85):

Definition 4.13. *(Provisional definition of stochastic integral for deterministic integrands)* If $f = f(t)$ is a function of t only, which is continuously differentiable, we put

$$\int_0^t f(s)dW_s := f(t)W_t - \int_0^t W_s f'(s)ds. \tag{89}$$

Remember here that $W_0 = 0$. The right hand side of (89) is well-defined, since $s \to W_s$ is continuous, and continuous functions can be integrated.

Does this settle (83)? Not quite, for what we are using there is an instance of the following general rule:

$$d\int_0^t f(s)dW_s = f(t)dW_t, \tag{90}$$

and to make things completely flawless we should check that (90) holds with the integral being *defined* by (89). Fortunately, this is easy: if we take the differential of the right hand side of (89), we find:

$$\begin{aligned} d\left(f(t)W_t - \int_0^t W_s f'(s)ds\right) &= W_t df(t) + f(t)dW_t - W_t f'(t)dt \\ &= f(t)dW_t, \end{aligned}$$

since $df(t) = f'(t)dt$. (**Attention!** We are very much using here that $f(t)$ is an ordinary, non-stochastic, function, and in particular that $df(t)$ *does not contain* a dW_t: see exercise 4.26 below for what could happen otherwise.)

We next compute the mean and variance of (89), these being interesting quantities for any rv. Put

$$I(t) = \int_0^t f(s)dW_s. \tag{91}$$

its expectation is, using (90):

$$\begin{aligned} \mathbb{E}(I_t) &= \mathbb{E}\left(f(t)W_t - \int_0^t W_s f'(s)ds\right) \\ &= f(t)\mathbb{E}(W_t) - \int_0^t f'(s)\mathbb{E}\left(W_s\right)ds \\ &= 0, \end{aligned} \tag{92}$$

where we used the linearity of the expectation to interchange it with the integral[12] and the fact that the mean of W_s, W_t is 0. Computing the variance is somewhat more involved, but the end result is a very simple and elegant formula:

Lemma 4.14. *(*First version of the Magical Formula of Stochastic Integration*) Let I_t be defined by (91). Then:*

$$\mathrm{Var}(I_t) = \mathbb{E}(I_t^2) = \int_0^t f(s)^2 ds. \tag{93}$$

We will leave the proof as an exercise with hints below, since later on we will derive a much more general result, for stochastic integrals in which the integrand $f(s)$ is allowed to be a stochastic process also, instead of just a deterministic function.

We can now finish our discussion of the solution of the Ornstein-Uehlenbeck equation:

[12] think of an integral as being a limit of Riemann sums!

Example 4.15. *(Ornstein-Uehlenbeck process, continued)* We recall that, by (84),

$$X_t = X_t = \theta(1 - e^{-\alpha t}) + X_0 e^{-\alpha t} + \sigma \int_0^t e^{\alpha(s-t)} dW_t.$$

is the solution of $dX_t = \alpha(X_t - \theta)dt + \sigma dW_t$. We can read off the mean and the variance of X_t from (92) and (93). First of all,

$$\mathbb{E}(X_t) = \theta(1 - e^{-\alpha t}) + X_0 e^{-\alpha t}. \tag{94}$$

Observe that $\mathbb{E}(X_t) \to \theta$, exponentially fast, as $t \to \infty$, regardless of the initial value X_0 we started with. The number θ therefore represents some kind of long-time equilibrium-value for the mean of X_t, which is independent of the initial value X_0 we started with. One uses the term "mean-reversion to θ" for this phenomenon. The parameter α indicates the speed with which this mean-reversion takes place. If for example (to fix ideas) we take $X_0 = 0$, then for $t = 1/\alpha$ we will have a relative error of:

$$\frac{|\mathbb{E}(X_{t=1/\alpha}) - \theta|}{|\theta|} = \frac{1}{e} \simeq .3679$$

Also note that if $\sigma = 0$, then the SDE is an ODE whose solution is simply (94) and we are done.

We next compute the variance of X_t:

$$\begin{aligned} \mathrm{Var}(X_t) &= \mathbb{E}\left((X_t - \mathbb{E}(X_t))^2\right) \\ &= \mathbb{E}\left(\left(\sigma \int_0^t e^{\alpha(s-t)} dW_t\right)^2\right) \\ &= \sigma^2 \int_0^t e^{2\alpha(s-t)} ds \quad \text{(by (93))} \\ &= \frac{\sigma^2}{2\alpha}\left((1 - e^{-2\alpha t}\right). \end{aligned} \tag{95}$$

We therefore see that X_t behaves, for big t, like a rv with mean approximately θ, fluctuating with a standard deviation of approximately $\sigma/\sqrt{2\alpha}$. Again, the influence of the initial value X_0 diminishes exponentially fast.

As a final remark we note that one can show that each X_t is actually a *normal* (or Gaussian) random variable. This is in fact true for any rv of the form

$$I_t = \int_0^t f(s) dW_s, \tag{96}$$

with a *deterministic* (that is, non-stochastic) f; see remark 4.16 below. Our final conclusion on the solution of the Ornstein-Uehlenbeck

equation is therefore that

$$X_t \sim N\left(\theta(1-e^{-\alpha t}) + X_0 e^{-\alpha t}, \frac{\sigma^2}{2\alpha}((1-e^{-2\alpha t})\right),$$

that is, each X_t is normally distributed, with the indicated mean and variance. For big $t > 0$ its pdf is approximately that of a $N(\theta, \sigma^2/2a)$-distribution.

One can go a bit further, and prove that the Ornstein-Uehlenbeck process is a Gaussian process, and compute the auto-covariances $\mathrm{cov}(X_s, X_t)$ (which, the process being Gaussian, completely determine it).

*__Remark 4.16.__ The normality of I_t can be understood as follows: think of I_t as a limit of Riemann sums: for $N \in \mathbb{N}$, put

$$s_j = \frac{j}{N}t,$$

so that $0 = s_0 < s_1 < \cdots < s_N = t$, and $|s_{j+1} - s_j| \to 0$ as $N \to \infty$ (properly speaking, the dependence of s_j on N should show up in the notations, but we won't do this, to keep the formulas simple). Then:

$$\begin{aligned}\int_0^t f(s)dW_s &:= f(t)W_t - \int_0^t W_s f'(s)ds \\ &= \lim_{N\to\infty} \sum_{j=0}^{N-1} W_(s_j) f'(s_j)\frac{t}{N}.\end{aligned}$$

Now $(W_{s_0}, W_{s_1}, \cdots, W_{s_{N-1}})$ is a normally distributed random vector, and $(f'(s_0), \cdots, f'(s_{N-1}))$ is an ordinary vector in $\mathbb{R}^N$. By general results for random vectors,

$$\sum_j f'(s_j)W_{s_j},$$

is normal (see the exercises at the end of this chapter), and it can be shown that a limit of normal rvs is again normal.

To make this last point completely rigorous we have to specify *what* we mean by saying that a sequence X_N, $N = 0, 1, 2, \cdots$ (in our case, the different Riemann sums) converge to a limiting rv Y. There are many different definitions possible, but it turns out that a convenient one is to say that $X_N \to Y$ if

$$\mathbb{E}\left((X_N - Y)^2\right) \to 0, \text{ as } N \to \infty.$$

This is called *mean square convergence*. One can show that mean square convergence implies that the mean of X_N tends to the mean of Y, and that the standard deviation of the difference $X_N - Y$ tends to 0. Moreover, one can show that limits in mean square of normal rvs will ne normal again; this is most easily using the so-called *characteristic functions* of the random variables involved.

4.7. **Multi-variable Ito calculus.** We begin by defining multi-dimensional Brownian motion. A *standard Brownian motion in* $\mathbb{R}^n$, or a *standard n-dimensional Brownian motion*, is simply a vector of n independent Brownian motions

$$\mathbb{Z}_t = (Z_{1,t}, \cdots, Z_{n,t}), \tag{97}$$

each $Z_{j,t}$ being a 1-dimensional Brownian motion (we suddenly use the letter Z instead of W since below we want to use the latter for a slightly more general process).

The Ito-rules for stochastic differentials are extended as follows:

$$dZ_{i,t}\, dZ_{j,t} = \delta_{ij} dt, \tag{98}$$

δ_{ij} being the Kronecker-delta. The reason is that, by independence, $dZ_{i,t}dZ_{j,t}$ has expectation 0, while its variance is $\mathbb{E}\left((dZ_{i,t})^2(dZ_{j,t})^2\right) = \mathbb{E}\left((dZ_{i,t})^2\right)\mathbb{E}\left((dZ_{j,t})^2\right) = dt \cdot dt = 0$.

A variant on this,which is often useful in Finance, is correlated Brownian motion. Let $\rho = (\rho_{ij})_{1\leq i,j\leq n}$ be a (constant) correlation matrix: ρ positive, $-1 \leq \rho_{i,j} \leq 1$ and $\rho_{ii} = 1$:

$$\rho = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \cdots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & 1 \end{pmatrix}, \tag{99}$$

and let $\mathbb{H} = (h_{ij})_{1\leq i,j\leq n}$ be an $n \times n$ matrix such that $\mathbb{H}\mathbb{H}^t = \rho$. If $\mathbb{Z}_t = (Z_{1.t}, \cdots, Z_{n,t})$ is a standard Brownian motion, we define $\mathbb{W}_t = (W_{1,t}, \cdots, W_{n,t})$ by:

$$\mathbb{W}_t = \mathbb{H}\mathbb{Z}_t, \tag{100}$$

or, in terms of components,

$$W_{i,t} = \sum_j h_{ij} Z_{j,t}. \tag{101}$$

This new process has the following, easily verified, properties:

- $\mathbb{W}_0 = 0$
- It $s \leq t$, then $\mathbb{W}_t - \mathbb{W}_s$ is multi-variate normal, with mean 0 and variance-covariance matrix $(t-s)\cdot\rho$:

$$\mathbb{W}_t - \mathbb{W}_s \sim N(0, (t-s)\cdot\rho).$$

- If $0 \leq u \leq s < t$, then $\mathbb{W}_u$ and $\mathbb{W}_t - \mathbb{W}_s$ are independent.

(Indeed, the first property is obvious, and as regards the second one, a linear combination jointly normal random vectors is again jointly normal, and a computation similar to one we did in section 3.4 shows

that

$$\begin{aligned}\mathbb{E}\left(\ (W_{i,t}-W_{i,s})(W_{j,t}-W_{j,s})\right) &= \sum_{k,l=1}^{n} h_{ik}h_{jl}\mathbb{E}\left((Z_{k,t}-Z_{k,s})(Z_{l,t}-Z_{l,s})\right)\\ &= \sum_{k=1}^{n} h_{ik}h_{jk}\mathbb{E}\left((Z_{k,t}-Z_{k,s})^2\right)\\ &= (t-s)\left(\mathbb{H}\mathbb{H}^t\right)_{ij}.\end{aligned}$$

Finally, the third property is an immediate consequence of the similar property of $\mathbb{Z}_t$.)

The process $(\mathbb{W}_t)_{t\geq 0}$ is called a *correlated Brownian motion*, with (constant) correlation matrix ρ. Note that if $\rho = Id$, the identity matrix, $\mathbb{W}_t$ is simply a standard Brownian motion (that is, its components are independent) since, for normal rvs, independence is equivalent with 0 correlation.

If $(\mathbb{W}_t)_{t\geq 0}$ is a correlated Brownian motion, then

$$dW_{i,t} = \sum_j h_{ij}dZ_{j,t}, \tag{102}$$

and, consequently,

$$\begin{aligned}dW_{i,t}\ dW_{k,t} &= \sum_j\sum_k h_{ij}h_{kl}dZ_{ij,t}dZ_{k,l}\\ &= \left(\sum_j h_{ij}h_{kj}\right)dt\\ &= \rho_{ik}dt,\end{aligned}$$

since $dZ_{j,t}\ dZ_{l,t} = \delta_{jl}$.

One can imagine having ρ depend on t: $\rho = \rho_t$ above. More generally, we could let $\mathbb{H}$ depend on t, without requiring that $\mathbb{H}\mathbb{H}^t$ is a correlation matrix, and add a t-dependent mean. This leads to the notion of multivariate Ito-process:

Definition 4.17. *(vector Ito process)* $(\mathbb{X}_t)_{t\geq 0}$ is called a *vector Ito process* if $d\mathbb{X}_t = (dX_{1,t}, \cdots .dX_{n,t})$ is related to an underlying n-dimensional standard Brownian motion $(\mathbb{Z}_t)_{t\geq 0}$ by:

$$dX_{i,t} = a_{i,t}dt + h_{ij,t}dZ_{j,t}, \tag{103}$$

where $a_{i,t}$ $(1 \leq i \leq n)$ and $h_{ij,t}$ $(1 \leq i,j \leq n$ are stochastic processes only depend on the Brownian motion through past up to present values:

$$a_{i,t},\ b_{ij,t} = \text{Functions of } \left((\mathbb{Z}_s)_{s\leq t}\right) \text{ and of } t. \tag{104}$$

$ $

In matrix notation, (103) can be written more concisely as:

$$d\mathbb{X}_t = a_t dt + \mathbb{H}_t d\mathbb{Z}_t, \tag{105}$$

where

$$a_t = \begin{pmatrix} a_{1,t} \\ \vdots \\ a_{n,t} \end{pmatrix}, \quad \mathbb{H}_t = \begin{pmatrix} h_{11,t} & \cdots & h_{1n,t} \\ \vdots & \ddots & \vdots \\ h_{n1,t} & \cdots & h_{nn,t} \end{pmatrix}.$$

With this terminology, a correlated Brownian motion is a special case of a vector Ito process, with an a_t which is 0, and a constant $\mathbb{H}_t$.

We next turn to Ito's lemma for vector-valued processes. Let $f = f(x,t)$ be a function on $\mathbb{R}^n \times \mathbb{R}$, where now $x = (x_1, \cdots, x_n) \in \mathbb{R}^n$.

Theorem 4.18. *(Ito's lemma for vector processes) Let $\mathbb{X}_t = (X_{1,t}, \cdots, X_{n,t})$ be a vector Ito process defined by (103). Then*

$$df(\mathbb{X}_t, t) = \frac{\partial f}{\partial t} dt + \sum_j \frac{\partial f}{\partial x_j} dX_{j,t} + \frac{1}{2} \sum_{j,k} \frac{\partial^2 f}{\partial x_j x_k} dX_{j,t} dX_{k,t}, \tag{106}$$

which can also be written as

$$\left(\frac{\partial f}{\partial t} + \sum_j a_{j,t} \frac{\partial f}{\partial x_j} + \frac{1}{2} \sum_{j,k,p} h_{jp,t} h_{kp,t} \frac{\partial^2 f}{\partial x_j x_k} \right) dt + \sum_j \sum_p h_{jp,t} \frac{\partial f}{\partial x_j} dZ_{p,t}. \tag{107}$$

Here,as before, all derivatives of f are to be evaluated in (X_t, t).

Proof. Despite the perhaps slightly formidable appearance of these formulas, their proof does not need any new ideas. One just has to know and apply the Taylor formula for an arbitrary number of variables: if $y = (y_1, \cdots, y_{n+1}) \in \mathbb{R}^{n+1}$ and $h \in \mathbb{R}^{n+1}$,

$$f(y+h) = f(y) + \sum_j \frac{\partial f}{\partial y_j} h_j + \frac{1}{2} \sum_{j,k} \frac{\partial^2 f}{\partial y_j y_k} h_j h_k + O(|h|^3),$$

apply this with $y = (x,t) = (x_1, \cdots, x_n, t)$ and $h = (dX_{1,t}, \cdots, dX_{n,t}, dt)$, and use the easily verified rules:

$$dX_{j,t}\, dt = (dX_{j,t})^3 = \cdots = 0.$$

This already proves (106) (which is the easiest to remember of the two formulas). Formula (107) then follows by substituting the expressions for $dX_{j,t}$ and observing that

$$\begin{aligned} dX_{j,t}\, dX_{k,t} &= \left(a_{j,t} dt + \sum h_{jp,t} dZ_{p,t} \right) \left(a_{k,t} dt + \sum_q h_{kq,t} dZ_{q,t} \right) \\ &= \sum_{p,q} h_{jp,t} h_{kq,t} dZ_{p,t} dZ_{q,t} \\ &= \sum_p h_{jp,t} h_{kp,t} dt, \end{aligned}$$

where we used (98). QED

Remark 4.19. Often, in financial models, one defines a process $\mathbb{X}_t$, in terms of a correlated Brownian motion $\mathbb{W}_t$ instead of $\mathbb{Z}_t$, the standard one. For example, the popular Heston stochastic volatility model is usually written as:

$$\begin{aligned} dS_t &= \mu S_t dt + \sqrt{v_t} dW_{1,t} \\ dv_t &= \alpha(\theta - \sigma_t)dt + \beta\sqrt{v_t} dW_{2,t}, \end{aligned} \tag{108}$$

where $(W_{1,t}, W_{2,t})$ are correlated Brownian motions with constant correlation ρ:

$$\mathbb{E}\left(dW_{1,t}\, dW_{2,t}\right) = \rho dt.$$

If we would want to evaluate $df(S_t, \sigma_t, t)$ (as one wants to in option pricing theory), we could first express $(W_{1,t}, W_{2,t})$ in terms of a standard 2-dimensional Brownian motion $(Z_{1,t}, Z_{2,t})$ (cf. exercise 4.31 below) and then apply formula (107) above, with $X_{1,t} = S_t$ and $X_{2,t} = \sigma_t$. However, in this situation it is much easier to start with the general form (106), and derive other forms computing directly with the $dW_{j,t}$. For example, we will now have

$$dX_{1,t}\, dX_{2,t} = \beta v_t S_t\, dW_{1,t} dW_{2,t} = \rho\beta v_t S_t dt.$$

More generally, suppose our Ito process $\mathbb{X}_t$ is given in terms of a correlated Brownian motion $\mathbb{W}_t$ with correlation matrix ρ:

$$d\mathbb{X}_t = a_t dt + \mathbb{H}_t d\mathbb{W}_t.$$

Since now $dW_{p,t} dW_{q,t} = \rho_{p,q}$, we now will find that $df(\mathbb{X}_t, t)$ equals:

$$\left(\frac{\partial f}{\partial t} + \sum_j a_{j,t}\frac{\partial f}{\partial x_j} + \frac{1}{2}\sum_{j,k}\sum_{p,q}\rho_{pq} h_{jp,t} h_{kp,t}\frac{\partial^2 f}{\partial x_j x_k}\right) dt + \sum_j \sum_p h_{jp,t}\frac{\partial f}{\partial x_j} dW_{p,t}.$$

Note that only the term involving the second derivatives of f as changed.

Finally, *vector SDE's* are simply vector Ito-processes for which the coefficients $a_{j,t}$ and $h_{jk,t}$ are functions of $\mathbb{X}_t$ itself and possibly of t:

$$d\mathbb{X}_t = a(\mathbb{X}_t)dt + \mathbb{H}(\mathbb{X}_t, t) d\mathbb{Z}_t, \tag{109}$$

or, written out in components:

$$\begin{aligned} dX_{1,t} &= a_{1,t}(\mathbb{X}_t, t)dt + \textstyle\sum_j h_{1j}(\mathbb{X}_t, t) dZ_{j,t} \\ &\vdots \\ dX_{n,t} &= a_{n,t}(\mathbb{X}_t, t)dt + \textstyle\sum_j h_{nj}(\mathbb{X}_t, t) dZ_{j,t} \end{aligned}$$

4.8. * **Sneak preview of general stochastic integration.** We end this (long) chapter by giving some idea of the general stochastic integral as it will be developed later on in this course, together with some motivation. The definition (89) has the drawback of only working for differentiable f, and it would be nice to be able to integrate a much larger class of f's, say continuous ones. Moreover, for many interesting financial applications we would like to replace the deterministic function $f(s)$ by a stochastic process, $(H_s)_{s\geq 0}$:

$$\int_0^t H_s dW_s, \tag{110}$$

and give a sense to integrals like

$$\int_0^t W_s dW_s, \int_0^t e^{W_s} dW_s,$$

(taking $H_s = W_s$ and e^{W_s}, respectively) or, more generally, to integrals of the form:

$$\int_0^t f(W_s, s) dW_s,$$

for a reasonably large class of functions f. Note that this would be completely hopeless with a definition like (89), due to the non-differentiability of Brownian motion. The key idea is to forget about (89), and return to a more primitive idea, the definition of integrals as limits of Riemann sums. One tries to make sense of (or, to be more precise, give a meaning to) the limit:

$$\int_0^t H_s dW_s = \lim_{N\to\infty} \sum_{j=0}^{N-1} H_{s_j} \left(W_{s_{j+1}} - W_{s_j} \right). \tag{111}$$

Note that the right hand side, as a limit of sums of rvs, will in general be a rv. The main mathematical problem is to show that the limit on the right exists in some sense. The correct interpretation will turn out to be that of mean square convergence (see remark 4.16* above) and for that to work two points will turn out to be be essential:

- If $u \leq s < t$, then H_u will have to be independent of $W_t - W_s$.
- The individual terms of the Riemann sum on the right are all evaluated "from the past to the future", in the sense that the integrand, H_{s_j}, is systematically evaluated in the left end point of the (stochastic) interval $[W_{s_j}, W_{s_{j+1}}]$.

Economically and financially, the two requirements make sense: suppose that W_t is the (future) price (per unit) of some asset or commodity, and H_t the number of units you (plan to) hold at time t; put otherwise, $(H_t)_{t\geq 0}$ is a *trading strategy*, which will in general be stochastic, since dependent on future, not yet realized, events. If you will trade only at the discrete times $s_0 = 0, s_1, \cdots, s_N$, then the total gain you will have made at t is precisely described by one of the sums on the right hand

side of (111): at time s_j, you hold H_{s_j}, which at time s_{j+1} will have given you a profit (or loss) of

$$H_{s_j}\left(W_{s_{j+1}} - W_{s_j}\right),$$

assuming there are no transaction costs. As a function of this profit/loss, you may then decide to change your holdings to $H_{s_{j+1}}$, etc. The condition of independence of H_u on future price changes $W_t - W_s$ is natural, since feasible trading strategies cannot depend on knowledge of not yet realized future price changes (which are, after all, unknown at the time of deciding your holding strategy).

If it is possible to trade continuously, and without transaction costs, its natural to take the limit, and write the total net profit at time t of the trading strategy $(H_t)_{t\geq 0}$ as a stochastic integral:

$$\int_0^t H_s dW_s.$$

(This above financial motivation of the stochastic integral has a defect, in the sense that W_t has a non-zero probability of being negative, which is somewhat unfortunate for a price. To be more realistic, we will need to give a sense to integrals like

$$\int_0^t H_\tau dS_\tau,$$

with $(S_t)_{t\geq 0}$ given by a geometric Brownian motion. This, however, turns out to be relatively easy, once we have understood integrals like (110)), for example since we have an explicit formula for S_t in terms of W_t.)

It will be shown that the notion of stochastic integral sketched here ties in nicely with the stochastic differentials we introduced before, in the sense that:

$$d\int_0^t f(W_s, s)dW_s = f(W_t, t)dW_t. \tag{112}$$

In fact, in the more rigorous mathematical treatment of the theory, one first defines stochastic integrals, and only later on introduces the stochastic differentials, via (112). We have reversed the logical order of things, on the assumption that the reader will feel that stochastic differentials are more intuitive to work with than stochastic integrals (the reader might of course disagree).

4.9. **Exercises to Chapter 4.**

Exercise 4.20. Use Ito's lemma to compute the stochastic differentials of the following functions of Brownian motion $(W_t)_{t\geq 0}$:
a) e^{W_t}.

b) $W_t^k, k \geq 0$.

c) $\cos(W_t)$.

d) $\arctan(t + W_t)$.

e) $e^{W_t^2}$.

f) $\cos(e^{W_t})$.

Exercise 4.21. Suppose that X_t is a stochastic process whose differential is given by:

$$dX_t = a(W_t, t)dt + \sigma(W_t, t)dW_t,$$

for given functions $a = a(w, t)$ and $\sigma = \sigma(w, t)$. If $f = f(x)$ is a twice differentiable function, derive an expression for $df(X_t)$ in terms of dW_t and dt, and use this to check your answers to parts e) and f) of the previous exercise.

Exercise 4.22. Find the solution of the following SDEs, a and σ being are arbitrary constants:

a) $dX_t = aX_t dt + \sigma dW_t$.

b) $dX_t = aX_t dt + \sigma X_t dW_t$.

c) $dX_t = aX_t(\theta - \log X_t)dt + \sigma X_t dW_t$.

Exercise 4.23. Find an integral expression for the solution of the following SDE:

$$dX_t = \alpha(\theta - X_t)dt + \sigma X_t dW_t.$$

(*Hint:* think part b) of the previous exercise, together with variation of constants.)

Exercise 4.24. Let $g = g(y)$ be a given function of y, and suppose that $y = f(w)$ is a solution of the ODE

$$dy = g(y)dw,$$

that is, $f'(w) = g(f(w))$. Show that $X_t = f(W_t)$ then is a solution of the SDE:

$$dX_t = \frac{1}{2}g(X_t)g'(X_t)dt + g(X_t)dW_t.$$

(*Hint:* Ito's lemma.)

Exercise 4.25. Solve the following SDE, and discuss up to which time the solution exists:

a) $dX_t = \sigma\sqrt{X_t}dW_t + \dfrac{\sigma^2}{4}dt$.

b) $dX_t = X^2 dW_t + X^3 dt$.

c) $dX_t = \cos^2 X_t - \dfrac{1}{2}\sin 2X_t \cos^2 X_t dt$.

d) $dX_t = X^k dW_t + \dfrac{k}{2}X^{2k-1}dt.$, k arbitrary.

e) $dX_t = e^{X_t} dW_t + \frac{1}{2} e^{2X_t} dt$.

Hint. Use the previous exercise and your knowledge in solving ODE's).

Exercise 4.26. Let X_t and Y_t be stochastic processes such that

$$dX_t = a_t dt + \sigma_t dW_t,$$

and

$$dY_t = b_t dt + \eta_t dW_t,$$

where $a_t, b_t, \sigma_t, \eta_t$ are given functions of t and W_t (this can be weakened). Show that

$$\begin{aligned} d(X_t Y_t) &= X_t dY_t + Y_t dX_t + dX_t dY_t \\ &= X_t dY_t + Y_t dX_t + \sigma(t)\eta(t) dt. \end{aligned} \tag{113}$$

(Observe that, again, a rule from ordinary calculus, the product rule, is augmented by an additional term, namely $dX_t dY_t$.)

Use this to show that if $f = f(t)$ is deterministic and differentiable, then

$$d(f(t)W_t) = f(t) dW_t + f'(t) W_t dt.$$

What does this become if f is a function of both time and Brownian motion, $f = f(W_t, t)$?

Hint for (113): $d(X_t Y_t) = X_{t+dt} Y_{t+dt} - X_t Y_t = (X_t + dX_t)(Y_t + dY_t) - X_t Y_t$, etc.) $$

Exercise 4.27. Use the method of variation of constants to find a solution of the SDE:

$$dX_t = \alpha X_t dt + (\gamma + \sigma X_t)\, dW_t.$$

Here α, γ and σ are constants.

*__Exercise 4.28.__ This exercise provides a proof of (93).
a) Show, by expanding the left hand side, that

$$\begin{aligned} \mathbb{E}(I_t^2) &= f(t)^2 \mathbb{E}(W_t^2) - 2 \int_0^t (f(t) f'(s) \mathbb{E}\,(W_t W_s ds) \\ &\quad + \int_0^t \int_0^t f'(s) f'(u) \mathbb{E}(W_s W_u)\, ds\, du. \end{aligned} \tag{114}$$

b) Show that the first term on the right hand side equals $t f(t)^2$.

c) Using that, for any s, t, we have that $\mathbb{E}(W_s W_t) = \min(s, t)$ (see exercise 3.11), show that the second term on the right in (114) is equal to:

$$2 \int_0^t s f'(s) ds = 2t f(t)^2 - 2 f(t) F(t),$$

where we have put

$$F(t) = \int_0^t f(s)ds,$$

a primitive of $f = f(t)$.

d) Verify the following set of identities for the last term in (114):

$$\begin{aligned} & \int_0^t \int_0^t \min(s,u) f'(s) f'(u) du ds \\ = \; & 2\int_0^t \left(\int_0^s u f'(s) f'(u) du \right) ds \\ = \; & 2\int_0^t s f'(s) f(s) ds - 2\int_0^t f'(s) \left(\int_0^s f(u) du \right) ds \\ = \; & t f(t)^2 - 2 f(t) F(t) + \int_0^t f(s)^2 ds. \end{aligned}$$

(*Hint:* For the final two equalities, just integrate by parts so many times that all derivatives $f'(s)$ have disappeared.)

e) Combine parts a) to d) to prove (93).

Exercise 4.29. Let X_t be the solution to the Ornstein-Uehlenbeck equation and suppose that $s < t$. Compute cov (X_s, X_t)

Exercise 4.30. Prove the following standard result on normally distributed random vectors: if $\mathbb{X} = (X_1, \cdots, X_N) \sim N(\mu, \mathbb{V})$ and if $v = (v_1, \cdots, v_N) \in \mathbb{R}^N$, and if $(v, \mathbb{X}) = v_1 X_1 + \cdots + v_N X_N$, then

$$(v, \mathbb{X}) \in N\left((v, \mu), (v, \mathbb{V}v)\right).$$

This fact was used in remark 4'16*.

Exercise 4.31. Let $-1 \leq \rho \leq 1$. a) Check that

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{pmatrix} \begin{pmatrix} 1 & \rho \\ 0 & \sqrt{1-\rho^2} \end{pmatrix}$$

b) Let $Z_{1,t}, Z_{2,t}$ be two independent Brownian motions. Explain why $(W_{1,t}, W_{2,t})$, defined by

$$\begin{aligned} W_{1,t} &= Z1, t \\ W_{2,t} &= \rho Z_{1,t} + \sqrt{1-\rho^2} Z_{2,t} \end{aligned},$$

is a correlated BM, with correlation matrix

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

c) Conversely, if $(W_{1,t}, W_{2,t})$ is as in b) with $\rho \neq \pm 1$, show that $W_{1,t}$ and $\widetilde{W}_{2,t}$, defined by

$$\widetilde{W}_t = \frac{1}{\sqrt{1-\rho^2}} \left(W_{2,t} - \rho W_{1,t} \right),$$

are two independent Brownian motions. What happens if $\rho = \pm 1$?

Exercise 4.32. Check that if $(W_t)_{t \geq 0}$ is a correlated Brownian motion, with positive definite (and therefore invertible!) correlation matrix ρ, and if $\mathbb{H}\mathbb{H}^t = \rho$, then $\mathbb{Z}_t := \mathbb{H}^{-1}\mathbb{W}_t$ is a standard Brownian motion.

5. Stochastic processes and PDE

There is a close connection between solving a SDE and solving boundary value problems for a certain type of partial differential equations (PDE) which are known as *parabolic*. Parabolic PDE's are roughly speaking those containing one time-derivative against two space derivatives. The precise definition will not be important to us; suffice to say that the majority of PDE's in finance, for example the Black and Scholes equation (cf. section 4.3 below), belong to this class.

Solutions to SDE are examples of an important class of stochastic processes which are called Markov processes, which we introduced in chapter 3. Under conditions on the coefficients $a = a(x,t)$ and $\sigma = \sigma(x,t)$ which guarantee the existence and uniqueness of a solution solutions of an SDE

$$dX_t = a(X_t,t)dt + \sigma(X_t,t)dW_t,$$

will be Markov. This is a non-trivial theorem of stochastic calculus, but intuitively speaking it is plausible: if we fix t and X_t, then the rv X_{t+dt} only depends on the value of X_t at time t: earlier values of X_t do not play a rôle.

An example of a non-Markov process would be one given by an equation like

$$dX_t = (X_t + X_{t/2})dW_t,$$

for which, in order to determine X_{t+dt}, we need to know both X_t and $X_{t/2}$.

5.1. **Chapman-Kolmogorov and Backwards Kolmogorov.**

Markov processes are uniquely determined by their transition probabilities

$$\mathbb{P}(X_s = y | X_t = x), \quad s > t,$$

but if we want to specify a Markov process we cannot choose these transition probabilities in a completely arbitrary way. They have to satisfy what are called the *Chapman-Kolmogorov equations*:

Theorem 5.1. *(Chapman-Kolmogorov) Let $(X_t)_{t\geq 0}$ be a Markov process, and let*

$$p(x,t;y,s) = \mathbb{P}(X_s = y | X_t = x), \quad t < s, \tag{115}$$

be its transition probability densities. Then for any time u between t and s, $t < u < s$, we will have that:

$$p(x,t;y,s) = \int_{\mathbb{R}} p(x,t;z,u)p(z,u;y,s)dz. \tag{116}$$

Proof: We first observe that if $t < u < s$,

$$\begin{aligned}&\mathbb{P}(X_s = y, X_t = x)\\ &= \int_{\mathbb{R}} \mathbb{P}(X_s = y, X_u = z, X_t = x)dz\\ &\int_{\mathbb{R}} \mathbb{P}(X_s = y|X_u = z)\mathbb{P}(X_u = z|X_t = x)\mathbb{P}(X_t = x)du,\end{aligned}$$

where we used the Markov property for the last line. If we divide both sides by $\mathbb{P}(X_t = x)$, and recall the definition of a conditional probability density and switch notations to (115) , we find (116). QED

The Chapman-Kolmogorov equations will now be used to prove an important connection between SDE's and PDE's:

Theorem 5.2. *Let $(X_t)_{t\geq 0}$ be a solution of the SDE*

$$dX_t = a(X_t, t)dt + \sigma(X_t, t)dW_t, \tag{117}$$

and let $f = f(x)$ be a given function. Fix a final time $T > 0$ and define a new function $V = V(x, t)$ for $t < T$ by:

$$V(x, t) = \mathbb{E}(f(X_T)|X_t = x). \tag{118}$$

Then $V = V(x, t)$ solves the following boundary value problem:

$$\begin{cases} \dfrac{\partial V}{\partial t} + \dfrac{\sigma(x,t)^2}{2}\dfrac{\partial^2 V}{\partial x^2} + a(x,t)\dfrac{\partial V}{\partial x} = 0\\ V(x, T) = f(x). \end{cases} \tag{119}$$

The partial differential equation for $V = V(x, t)$ is called Kolmogorov's backward equation *associated to the SDE (117).*

Proof. We first observe that, by the definition of expectation,

$$V(x, t) = \int_{\mathbb{R}} p(x, t; y, T)f(y)dy. \tag{120}$$

Next, if we use the Chapman-Kolmogorov integral equation (116) with $u = t + dt$, we find that

$$\begin{aligned} V(x, t) &= \int_{\mathbb{R}}\int_{\mathbb{R}} p(x, t; z, t + dt)p(z, t + dt; y, T)f(y)dydz\\ &= \int_{\mathbb{R}} p(x, t; z, t + dt)V(z, t + dt)dz \end{aligned}$$

so that

$$V(x, t) = \mathbb{E}\left(V(X_{t+dt}, t + dt)|X_t = x\right). \tag{121}$$

Next, by Ito's lemma (V can be sown to be twice continuously differentiable),

$$\begin{aligned} V(X_{t+dt}) &= V(X_t) + \frac{\partial V}{\partial x}(X_t,t)dX_t + \frac{1}{2}\frac{\partial^2 V}{\partial x^2}(X_t,t)(dX_t)^2 \\ &= V(X_t) + \left(a(X_t,t)\frac{\partial V}{\partial x}(X_t,t) + \frac{\sigma(X_t,t)^2}{2}\frac{\partial^2 V}{\partial x^2}(X_t,t)\right)dt \\ &\quad +\sigma(X_t,t)\frac{\partial V}{\partial x}(X_t,t)dW_t. \end{aligned}$$

Taking conditional expectations, and observing that

$$\mathbb{E}(\sigma(X_t,t)dW_t|X_t=x) = \sigma(x,t)\mathbb{E}(dW_t) = 0,$$

we find that

$$\mathbb{E}\left(V(X_{t+dt},t+dt)|X_t=x\right) = V(x,t)+a(x,t)\frac{\partial V}{\partial x}(x,t)+\frac{\sigma(x,t)^2}{2}\frac{\partial^2 V}{\partial x^2}(x,t).$$

Substituting this in (121), we find that $V=V(x,t)$ satisfies the stated differential equation.

Finally, for the boundary condition at $t=T$, it is obvious that $\mathbb{E}(f(X_T)|X_T=x) = f(x)$. QED

For applications in Finance the following extension of theorem 5.2 is often important.

Theorem 5.3. *(Baby Feynman-Kac) Keeping the notations of the previous theorem, let $\rho=\rho(t)$ be a deterministic function of time t, and define*

$$V(x,t) = e^{-\int_t^T \rho(s)ds}\mathbb{E}(f(X_T)|X_t=x). \tag{122}$$

Then $V(x,t)$ solves the PDE

$$\frac{\partial V}{\partial t} + \frac{\sigma(x,t)^2}{2}\frac{\partial^2 V}{\partial x^2} + a(x,t)\frac{\partial V}{\partial x} = \rho(t)V, \tag{123}$$

with the same boundary condition as before: $V(x,T)=f(x)$.

Proof. The proof is easy: simply observe that

$$\widetilde{V}(x,t) = e^{\int_t^T \rho(s)ds}V(x,t) = \mathbb{E}(f(X_T)|X_t=x),$$

satisfies the PDE of theorem 5.2:

$$\frac{\partial \widetilde{V}}{\partial t} + \frac{\sigma(x,t)^2}{2}\frac{\partial^2 \widetilde{V}}{\partial x^2} + a(x,t)\frac{\partial \widetilde{V}}{\partial x} = 0,$$

substitute $\widetilde{V} = V\exp(\int_t^T \rho(s)ds)$, and observe that

$$\frac{\partial \widetilde{V}}{\partial t} = \left(\frac{\partial V}{\partial t} - \rho(t)\right)e^{\int_t^T \rho(s)ds}.$$

QED

There is a generalization of theorem 5.3 for when ρ is also allowed to depend on x:

$$\rho = \rho(x,t).$$

In this case the solution to the boundary value problem is given by the famous Feynman-Kac formula:

Theorem 5.4. *(Feynman-Kac) Let*

$$V(x,t) = \mathbb{E}\left(e^{-\int_t^T (\rho(X)s,s)ds} f(X_T)|X_t = x\right). \tag{124}$$

Then $V = V(x,t)$ solves

$$\frac{\partial V}{\partial t} + \frac{\sigma(x,t)^2}{2}\frac{\partial^2 V}{\partial x^2} + a(x,t)\frac{\partial V}{\partial x} = \rho(x,t)V,$$

and satisfies the same boundary condition as before, $V(x,T) = f(x)$.

*__Remark 5.5.__ This theorem cannot be proved anymore by a simple trick such as that one used for theorem 5.3. Rather, one has to repeat the proof of theorem 5.2. A problem which occurs is that, due to the presence of the term $\int_t^T V(X_s,s)ds$ on the right in (124), we cannot express $V(x,t)$ anymore by a simple integral involving a single transition transition probability , as we could in equation (120): we now need to take into account the transition probabilities at infinitely (even continuously) many intermediary times $t < s < T$. Trying to do this by brute force this[13], leads to complicated and rather messy formulas. As we will see later on, it is possible to give a rather slick proof along the same lines as the one of 5.2, after we will have developed the machinery of measure-theoretic probability.

5.2. **Solving the Black and Scholes PDE with probability.** You may already be acquainted with the Black and Scholes equation, which states that any option written on an underlying stock whose stock-price follows a geometric Brownian motion $dS_t = \mu S_t dt + \sigma S_t dW_t$ has a price $V = V(S,t)$ which, *before exercise*, satisfies the following PDE:

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S_t^2 \frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} = rV; \tag{125}$$

Here $V(S,t)$ should be read as "the price of the option at time t if $S_t = S$.", and r is the (constant) interest rate. The Black and Scholes equation will be explained in detail in the Pricing module of this MSc. European options can only be exercised at a time of maturity, T, and their final pay-off provides a boundary condition at $t = T$. For example, for a call:

$$V_{\text{Call}}(S,T) = \max(S - E, 0),$$

[13]Basically, by approximating the integral by Riemann sums, for which one only has to take into account the large, but finite, number of times s_j which we use to partition the interval, and then take the limit over all Riemann sums. This was Feynmann's original approach, except that he did this for another equation, the Schrödinger equation from Quantum Mechanics

and, for a general European pay-off of $f(S_T)$ at T:

$$V(S,T) = f(S), \tag{126}$$

f being specified by the derivative's contract.

Now (125), (126) is precisely the type of problem we can apply our Baby Feynman-Kac theorem, theorem 5.3, to: if we use s instead of t as a time-variable instead of t (since t now is the, fixed, time at which we wish to evaluate $V(S,t)$), and if $(X_s)_s$ solves:

$$dX_s = rX_s ds + \sigma X_s dW_s, \tag{127}$$

then

$$V(S,t) = e^{-r(T-t)}\mathbb{E}\left(f(X_T)|X_t = S\right). \tag{128}$$

Since

$$X_s = Se^{(r-\sigma^2/2)(s-t)+\sigma(W_s-W_t)}$$

solves (127) and is equal to S at $s = t$, we find that, writing $\tau = T - t$ for the time-to-maturity,

$$\begin{aligned} e^{r(T-t)}V(S,t) &= \mathbb{E}\left(f(Se^{(r-\sigma^2/2)(T-t)+\sigma(W_T-W_t)})\right) \\ &= \int_{\mathbb{R}} f\left(Se^{(r-\sigma^2/2)\tau+\sigma w}\right) e^{-w^2/2\tau}\frac{dw}{\sqrt{2\pi\tau}}, \end{aligned}$$

since $W_T - W_t \sim N(0,\tau)$. The integral can be evaluated for many pay-offs. If we take $f(S) = \max(S - E, 0)$, one finds the celebrated Black and Scholes formula for the price of an European call: the necessary calculations are the same as in (extra) exercise 2.32.

Note that the auxiliary process X_t we introduced to solve the PDE follows a geometric Brownian motion, just like the price process for S_t, but with the difference that the expected return is r instead of μ. It is like the price of an (artificial) security which has a risk-less rate of return, r, although there is a price risk present, in the form of the Brownian motion term, assuming that $\sigma > 0$ (if $\sigma = 0$, there is no risk: the price evolves deterministically, not stochastically). The only kind of investor who would invest in such a security would be one who doesn't care about risk, in that he does not require an additional return $\mu - r$ as a reward for taking risk. Such (hypothetical) investors are called risk-neutral, and (128) is called the *discounted* (due to the $e^{-r(T-t)}$ in front) *risk-neutral expectation* of the final pay-off $f(S_T)$. Pricing formulas in Finance for derivative assets with European exercise often ave this structure:

$$(\text{Value at } t\ ) = \mathbb{E}_{\text{risk-neutral}}\left((\text{discount factor})\cdot(\text{pay-off})\ |S_t\ \right),$$

where the discount factor may in general be stochastic also, and is therefore put inside the expectation-sign. The connection between a PDE for a price, and such an expectations-type formula is always given by a Feynmann-Kac theorem.

The derivation of the Black and Scholes PDE shows why we are led to a "risk-neutral process" X_t instead of the "risk-rewarding one ", S_t: the risk is hedged away, and the growth rate of the stock does not play a rôle in the pricing of the derivative. It is also possible to derive (128) directly, without passing via PDE's, by what is called the martingale pricing method. This will be explained in the Spring semester of the Pricing module.

5.3. **Exercises to chapter 5.**

Exercise 5.6. Consider the boundary value problem for the backwards heat equation:

$$\begin{array}{l} \dfrac{\partial V}{\partial t} + \dfrac{1}{2}\dfrac{\partial^2 V}{\partial x^2} = 0, \\ V(x,1) = f(x), \end{array}$$

where $f(x)$, the boundary value for time $t = 1$, is given, and where we are looking for a function $V = V(x,t)$ defined for $t \leq 1$ and $x \in \mathbb{R}$. Use theorem 5.2 to show that a solution is given by the following formula:

$$V(x,t) = \int_{\mathbb{R}} f(y) e^{-(x-y)^2/2(1-t)} \frac{dy}{\sqrt{2\pi(1-t)}}.$$

One has to worry a bit about for which f's the formula makes sense. Can you think of an f for which it wouldn't?

Exercise 5.7. Now consider the boundary problem for the heat equation with a drift term:

$$\begin{array}{l} \dfrac{\partial V}{\partial t} + \dfrac{1}{2}\dfrac{\partial^2 V}{\partial x^2} + a\dfrac{\partial V}{\partial x} = 0, \\ V(x,1) = f(x), \end{array}$$

a being a constant. Derive an explicit integral formula for the solution $V = V(x,t)$, along the lines of the previous exercise.

Exercise 5.8. A *European digital call* is an option which, at maturity T, will pay you 1 if the stock price S_T at T is bigger than the exercise price E, and 0 otherwise. In terms of the Heaviside function $H = H(x)$, defined by

$$H(x) = \begin{cases} 1 \text{ if } x > 0, \\ 0 \text{ if } x \leq 0, \end{cases}$$

the pay-off of such a digital call equals $H(S_T - E)$. Suppose that the price of the derivative satisfies the Black-Scholes equation before exercise.

a) Derive an explicit formula for the price of a European digital call.

b) Derive a formula for the value of a European digital put, with pay-off $H(E - S_T)$.
(*Hint:* derive a put-call parity relationship for digital calls and puts; what do you notice about $H(S_T - E) + H(E - S_T)$?)

Exercise 5.9. Derive the Black and Scholes formula for a European call.

Exercise 5.10. Consider the following boundary value problem:

$$\begin{cases} \dfrac{\partial V}{\partial t} + \alpha\,(\theta - x)\,\dfrac{\partial V}{\partial x} + \dfrac{\sigma^2}{2}\dfrac{\partial^2 V}{\partial x^2} = 0, \quad \text{for } t < T\ , \\ V(x,T) = g(x), \end{cases}$$

where $g = g(x)$ is some given function. What is the stochastic process associated to this PDE by Kolmogorov's theorem 5.2? Write the solution to the boundary value problem as an expectation.

6. Measure-theoretic probability

In this chapter we will try to familiarize you with the language of measure theoretic probability theory. A probability will be seen as a map giving different weights, between 0 and 1, to possible future events our "outcomes of statistical experiments", the latter being seen as sets of "potential future states of the world" compatible with the event (or experimental outcome) in question. To formalize these notions we will make heavy use of set-theoretical notation; cf. the additional handout.

In brief, to do probability, we will need a

- *sample space* Ω,
- a *σ-algebra* $\mathcal{F}$,
- a *probability measure* $\mathbb{P}$, and
- the idea of a rv as a real-valued function on Ω which is measurable with respect to $\mathcal{F}$.

We will examine these concepts one by one, in separate subsections of this chapter. We will abundantly use set-theoretic notation, which is briefly reviewed in the final subsection to this chapter, for convenience of the reader.

6.1. **Sample spaces.** The sample space Ω is a (usually very big) set, which can be thought of as the set of all possible future states of the world, or as the set of all possible outcomes of some statistical experiment. Simple examples are:

- Tossing a coin once. Then $\Omega = \{H, T\}$, where "H " stands for "Heads" and "T" for "Tails".
- Tossing a coin thrice. The sample space now consists of all three-term sequences which we can form out of H and T:

$$\Omega = \begin{array}{l}\{HHH, HHT, HTH, HHT,\\ HTT, THT, TTH, HHH\}.\end{array} \tag{129}$$

  Each of the elements of Ω represents a possible outcome of this coin-tossing experiment: three Heads in a row, two Heads in a row and then one tail, a Head, a Tail and then a Head, etc.
- Tossing a coin infinitely many times: Ω will now consist of all possible infinite sequences of Heads and Tails

$$(HTHHTHHTH\cdots),$$

  etc. More formally:

$$\Omega = \{(x_1x_2x_3\cdots) : x_j = H \text{ or } x_j = T\}.$$

  You may not realize this at first, but this is a *huge* set: if we would let x_j be 0 or 1 above, instead of H or T, then such a

sequence $(x_1, x_2, \cdots)$ can be thought of as a real number between 0 and 1, written in base 2 (or in bits, of you like):

$$x_1 2^{-1} + x_2 2^{-2} + x_3 2^{-3} + \cdots ,$$

and there are therefore as many elements of Ω as there are real numbers between 0 and 1.

Finally, two very important examples for finance:

- *Observing the price of an asset S at discrete times $t = 0, 1, 2, \cdots$.* In this case, Ω is the set of all infinite sequences

$$(s_0, s_1, s_2, \cdots), \ s_j \in \mathbb{R}_{\geq 0}, \text{ for all } j \ .$$

where s_j = price of P at $t = j$ "in the state of the world $\omega = (s_0, s_1, s_2, \cdots)$".

Mathematically speaking, such a sequence can be thought of as a function from $\mathbb{N}$ to the set of positive reals, $[0, \infty)$ (sometimes also denoted by $\mathbb{R}_{\geq 0}$).

- *Observing an asset price continuously, or in continuous time.* In this case we can take Ω to be the set of all *all functions from the positive reals to the positive reals*:

$$\Omega = \{s : s : [0, \infty) \to [0, \infty)\}.$$

The asset's price at time t is $s(t)$ (the value of s at t), if we are in the state of the world $\omega = s \in \Omega$. In other words, Ω is simply the set of all possible price trajectories for our stock S.

In this example we often use smaller sets of functions, for example by requiring that $s : [0, \infty) \to [0, \infty)$ is continuous. If we allow our prices to jump, a convenient space is that of functions which are continuous to the right with a left limit, or RCLL. Another acronym for such functions in the mathematical literature is "càdlàg", from "continue à droite, limites à gauge".

6.2. **The σ-algebra of events.** Typically we do not know which state the world is in: referring to the above examples, we do not know in advance how future prices of a given stock will develop, or what the total outcome of a repeated coin tossing element will be. We seek knowledge by doing observations or, statisticians would like to say, by performing experiments, for example looking up today's price of your favorite stock in the FT. Such an observation will not, in general, completely determine the state of the world we are inc, but only partially. E.g., you might consult your FT only every other day, and therefore remain ignorant of what the stock's price has been in between. *Events* are subsets $F \subset \Omega$ which can be thought of as sets of all possible states of the world (that is, points in sample space) which are compatible with a given (hypothetical) experiment or observation. For example,

- In the continuous time asset price example above, the event F might be:

$$F = \{s : [0, \infty) \to [0, \infty) : s(t_0) > 1\},$$

the set of all possible price trajectories for which the price at a given time t_0 will exceed 1. More complicated events easily suggest themselves, for example

$$F = \{s : [0, \infty) \to [0, \infty) : \frac{s_{t_0} - s_{t_0-h}}{s_{t_0-h}} > 0.05\},$$

the event that at t_0 the stock's return over $[t_0 - h, t_0]$ will have exceeded 5%.
- Another, more elementary, example is, in the second dice-throwing example above, the event that the third throw will give a Head. This is represented by the subset

$$F = \{HHH, HTH, THH, TTH\},$$

of Ω.

Being subsets of Ω, we can consider unions, intersections and complements of events. This will lead to an "algebra of events". More specifically, if F and G are events, then we can form:

(1) Their intersection,

$$F \cap G = \{\omega \in \Omega : \omega \in F \text{ and } \omega \in G\}.$$

In the statistical context of making observations (doing a statistical experiment) this is interpreted as:

- $F \cap G$: the event that both F and G will occur.

(2) Their union,

$$F \cup G = \{\omega \in \Omega : \omega \in F \text{ or } \omega \in G\},$$

interpreted as:

- $F \cup G$: the event that either F or G will happen.

It is important to realize that the "or" is not exclusive: it is allowed that F and G happen simultaneously.

(3) The complement of F in Ω, $\Omega \setminus F$, defined by

$$\Omega \setminus F = \{\omega \in \Omega : \omega \notin F\},$$

where $\notin$ means: "not an element of". The interpretation is that

- $\Omega \setminus F$ is the event that the event F will not occur.

To save time, we often write F^c for the complement: $F^c := \Omega \setminus F$.

We also single out two very special events, namely the entire set Ω itself, corresponding to "something will always happen" (but you're not at all interested in what), and the empty set $\emptyset$, the set with no elements,

corresponding to "something impossible will occur" (e.g. $F \cap (\Omega \setminus F)$).

As the next step, we will now consider *collections* of events $\mathcal{F}$sets of events $\mathcal{F}$ which contain both $\emptyset$ and Ω and which are closed under operations (1) to (3). We will in fact impose a stronger version of (2):

Definition 6.1. *(σ-algebra).* A σ-algebra on Ω is a set $\mathcal{F}$ of subsets of Ω such that:

(i) Both $\emptyset \in \mathcal{F}$ and $\Omega \in \mathcal{F}$.

(ii) If $F_1, F_2, \cdots, F_n, \cdots \in \mathcal{F}$, then their (countably infinite) union is also in $\mathcal{F}$:

$$\cup_{n=1}^{\infty} F_n \in \mathcal{F}.$$

(iii) If $F \in \mathcal{F}$, then also $F^c = \Omega \setminus F \in \mathcal{F}$.

Note that individual elements of $\mathcal{F}$ are themselves sets, namely subsets of Ω.

If $\mathcal{F}$ only satisfies (ii) for *finite* unions (but still satisfies (i) and (iii)) it is called an *algebra* instead of a σ-algebra[14].

The attentive reader will have noticed that we did not explicitly require that intersections are in $\mathcal{F}$, but this is in fact a consequence of the other conditions (i) and (ii): for example, it is easily checked that

$$(F \cap G)^c = F^c \cup G^c,$$

and therefore

$$F \cap G = \Omega \setminus (F^c \cup G^c) = (F^c \cup G^c)^c,$$

is in $\mathcal{F}$, if F and G are. This extends to infinite intersections: see the exercises.

There exist two somewhat special examples of σ-algebras:

- The *trivial* σ-algebra: $\mathcal{F}_{triv} = \{\emptyset, \Omega\}$.
- The σ-algebra consisting of *all* possible subsets of Ω, which is also called the powerset of Ω; we will call it the *discrete $\sigma-$algebra:*

$$\mathcal{F}_{discr} = \mathcal{F}_{discr}(\Omega). \tag{130}$$

This latter σ-algebra is in most cases the appropriate one if Ω consists of a finite set, like in the finite coin-tossing experiments above, but is in general much too big to be useful when Ω is infinite.

Before looking at some more involved (and more interesting) examples we have to explain a way of generating σ-algebras. Suppose we

[14]The suffix "σ" is often used in mathematics theory to indicate countably infinite unions or sums

start off with some collection $\mathcal{C}$ of subsets of Ω which does not necessarily satisfy either one of the conditions (i), (ii) and (iii) of definition 6.1.We now enlarge this initial set $\mathcal{C}$ by throwing in arbitrary (countable) unions of elements of $\mathcal{C}$, complements of elements of $\mathcal{C}$, and then continue by adding arbitrary (countable) unions of complements of elements of $\mathcal{C}$, complements of countable unions of elements of $\mathcal{C}$, and so on, continuing until do not get ant new sets anymore to add. It is intuitively clear that in this way we will end up with a σ-algebra which is called the *σ-algebra generated by $\mathcal{C}$*, and denoted by

$$\sigma(\mathcal{C}). \tag{131}$$

Armed with the concept of σ-algebras generated by sets of subsets, we can now give some more ambitious examples of the former.

*__Remark 6.2.__ What we just have given amounts to a kind of "bottom up" description of $\sigma(\mathcal{C})$. It is a bit inconclusive in that $\sigma(\mathcal{C})$ appears as the limiting result of a never-ending set of operations of taking unions and complements. The usual mathematical definition is more "top down", and defines $\sigma(\mathcal{C})$ as the intersection of all σ-algebras containing $\mathcal{C}$:

$$\sigma(\mathcal{C}) = \cap\{\mathcal{F} : \mathcal{F} \ \sigma\text{-algebra}, \ \mathcal{C} \subset \mathcal{F}\}.$$

This corresponds to thinking of $\sigma(\mathcal{C})$ as being the *smallest σ-algebra containing $\mathcal{C}$*. To make it work one has to check that the intersection of a collection of σ-algebras (here, that of all those containing $\mathcal{C}$) is again a σ-algebra; we leave the verification of this fact to the reader. If this is too abstract to your taste, you may prefer the informal description above.

Example and Definition 6.3. An important example of this construction is given by the σ-algebra of *Borel subsets of* $\mathbb{R}$. Here we take for $\mathcal{C}$ just the set of all open intervals (a, b) of $\mathbb{R}$. The σ-algebra generated by this $\mathcal{C}$ is called the *Borel σ-algebra of* $\mathbb{R}$*:*

$$\mathcal{B}(\mathbb{R}) = \sigma\left(\{(a, b) : a < b\}\right).$$

We can do something similar in $\mathbb{R}^n$: here the basic building blocks we start from open rectangles, which are simply products of open intervals:

$$\begin{aligned} R &= (a_1, b_1) \times (a_2, b_2) \times \cdots \times (a_n, b_n) \\ &= \{x = (x_1, \cdots, x_n) : a_j < x_j < b_j\}, \end{aligned}$$

and we define $\mathcal{B}(\mathbb{R}^n)$ to be the σ-algebra generated by the set of all such rectangles.

Example 6.4. *(σ-algebra of the infinite coin-tossing experiment)* Recall that

$$\Omega = \{\omega = (x_1 x_2 x_2 \cdots) : x_j = H \text{ or } T\}.$$

In practice, we will only be able to observe a finite part of such an experiment, and will only be able to decide whether a segment of an

element $\omega \in \Omega$ (which is an infinite sequence of Heads and Tails) coincides with some pre-determined pattern of H's and T's. This suggest to consider the following type of events: let $y_1 \cdots y_n$ be some sequence of Heads and Tails, $y_j = T$ or H, for each j, which we should think of as given beforehand. The following subset of Ω will then correspond to the event that, in a coin-tossing experiment, an infinite sequence of tosses will start with $y_1 y_2 \cdots y_N$:

(132)
$$F_{y_1 \cdots y_N} = \{\omega = (x_1 x_2 x_2 \cdots) \in \Omega : x_1 = y_1, x_2 = y_2, \cdots, x_N = y_N\}.$$

For example, F_{HHT} is the event that the tossing experiment will begin with a "Head, Head, Tails".

The natural σ-algebra is the one generated by the events (132), with arbitrary $N \in \mathbb{N}$ and arbitrary $y_j \in \{H, T\}$.

The following two examples concern financial applications:

Example 6.5. *(Natural σ-algebra for asset prices in discrete time)* See under (i) for the definition. Again, we will only be able to observe at at most a finite set of times $t_1 = n_1, \cdots, t_k = n_k$, where $n_1, \cdots, n_k$ are given natural numbers. At each of these times t_j, a typical observation will be whether the stocks price is in some given interval (or "window") I_j, e.g. $I_j = (a_j, b_j)$. Such an observation corresponds to the event

$$F_{t_j, I_j} = \{(s_0, s_{1,2}, \cdots) : s_{n_1} \in I_1, s_{n_2} \in I_2, \cdots, s_{n_k} \in I_k\} \tag{133}$$
$$= \{s = (s_0, s_{1,2}, \cdots) : a_1 < s_{N_1} < b_1, \cdots, a_k < s_{n_k} < b_k\}.$$

For example, we might consider the event that the price, 2 days from now, is between 5 and 7, and that in 4 weeks it will risen above 12:

$$\{s = (s_1, s_2, \cdots) : 5 \leq s_2 \leq 7, s_{28} \geq 12\}.$$

The natural σ-algebra now is the one generated by all sets of the form (133). One can check that this σ-algebra consists precisely of all sets of the form:

$$\{(s_1 s_2 \cdots) : s_1 \in B_1, s_2 \in B_2, \cdots, s_n \in B_n, \cdots\},$$

where $B_1, B_2, \cdots$ is a sequence of Borel sets which ly in $[0, \infty)$.

Example 6.6. *(Natural σ-algebra for asset prices in continuous time)* In this case we had

$$\Omega = \{s : [0, \infty) \to [0, \infty) \text{ function}\},$$

and again, the basic type of observation we can do is to see whether for some finite set of times $t_1, \cdots t_N$, the asset prices is or is not in certain intervals I_j. This corresponds to the event:

$$F_{(t_j)_j, (I_j)_j)} = \{s : [0, \infty) \to [0, \infty) : s(t_j) \in I_j\}, \tag{134}$$

and the natural σ-algebra is the one generated by all of such events, for all possible N and all possible times t_j and intervals I_j $(1 \leq j <\leq N)$.

Again, one can show that this σ-algebra consists precisely of all sets of functions of the following form:

$$F_{(t_j)_j,(B_j)_j} = \{s : [0,\infty) \to [0,\infty) : s(t_j) \in B_j, \text{ for all } j \in \mathbb{N}\}, \tag{135}$$

$(t_0, t_1, \cdots)$ being an, in principle infinite, sequence of times, and $B_j \in \mathcal{B}(\mathbb{R})$ a sequence of Borel subsets of $[0,\infty)$.

*__Side-remark 6.7.__ *(can be ignored without loss of continuity)* Note that in these examples, the countable "union axiom" (ii) naturally leads to events which require observations at infinite set of times $t_0, t_1, \cdots$. A similar remark applies to the condition of σ-additivity in the definition of a probability, cf. definition 6.8 below. This is of course not realistic, and should be considered a mathematical idealization, whose final justification is to be found in the smooth and flexible mathematical theory which results from in. Indeed, it is possible to develop a probability (or measure) theory based on algebras, and on finitely additive measures, but this theory actually turns out to be more complicated. Interestingly enough, these finitely additive measures (or "charges", as they are sometimes called) do play a role in certain very recent work on optimization problems in finance, where they, so to speak, enter again through a back-door, basically since they play a role in the description of the dual space of the set of bounded functions on a probability space.

6.3. **Probability measures.** We next introduce probabilities in all of this. A *probability* $\mathbb{P}$ assigns to each event $F \in \mathcal{F}$ a number $\mathbb{P}(F)$ between 0 and 1, interpreted as the *probability that F will occur in an experiment or observation*, in such a way that certain obvious requirements are met.

Definition 6.8. Given a sample space Ω, and a σ-algebra of events $\mathcal{F}$ on Ω, a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ is a function

$$\mathbb{P} : \mathcal{F} \to [0,1],$$

such that the following two conditions are satisfied:

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

(b) (σ-additivity) If $F_1, F_2, \cdots, F_j, \cdots$ is a, possibly infinite, sequence of events $F_j \subset \mathcal{F}$ which are *mutually exclusive*, in the sense that

$$F_j \cap F_k = \emptyset \text{ if } j \neq k,$$

then

$$\mathbb{P}\left(\cup_j F_j\right) = \sum_j \mathbb{P}(F_j).$$

Interpretation: (a) just states that, with probability 1 , "something will occur". As for (b), mutually exclusive events are those which cannot be observed simultaneously (like a stock's price t time t_1 being both > 2 and < 1), and for those, the probability that one of these will

happen is just the sum of the individual probabilities. As indicated, condition (b) is called σ-*additivity* of $\mathbb{P}$.

If we drop condition (a), and allow that $\mathbb{P}$ takes on arbitrary positive values, we obtain the definition of a measure.

Definition 6.9. A *measure* on $\mathcal{F}$ is a map

$$\mu : \mathcal{F} \to [0, \infty],$$

which satisfies condition (b) of definition 6.8 (with $\mathbb{P}$ of course replaced by μ.

One should think of a measure as being an abstraction, or generalization, of the familiar concepts of length, surface area, or volume. Their main point is given a measure, we can integrate functions, as we will explain below.

Simple consequences of definition 6.8. Before turning to examples we first look at some simple consequences of definition 6.8. First, as is to be expected, the probability of $F \in \mathcal{F}$ not happening is $1 - \mathbb{P}(F)$:

$$\text{If } F \in \mathcal{F}, \text{ then } \mathbb{P}(\Omega \setminus F) = 1 - \mathbb{P}(F). \tag{136}$$

For this it suffices to apply (b) of definition 6.8 to $F_1 = F$ and $F_2 = \Omega \setminus F$ (which are of course mutually exclusive, by definition).

If we take in particular $F = \Omega$ in (136), we find that

$$\mathbb{P}(\emptyset) = 0, \tag{137}$$

since $\emptyset = \Omega \setminus \Omega$.

Example 6.10. *(Coin tossing again)* If we toss a coin once, then $\Omega = \{H, T\}$, and we can take $\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$. We can define a probability measure $\mathbb{P}$ by declaring:

$$\mathbb{P}(\{H\}) = \mathbb{P}(\{T\}) = \frac{1}{2}.$$

(This would be a fair coin; how would you model a coin where the probability of Heads is, for example, twice the probability of Tails?).

Observe that $\mathbb{P}(\{H, T\}) = \mathbb{P}(\Omega) = 1 = \mathbb{P}(\{H\}) + \mathbb{P}(\{T\})$, so that (b) of definition 6.8 is satisfied.

Now let us toss the same coin thrice, and use the Ω of (129), and the σ-algebra $\mathcal{F}_{discr}$ consisting of all possible subsets of Ω. It is important to realize that to define a probability measure $\mathbb{P}$, it suffices to prescribe the values of $\mathbb{P}$ on each of the elements of Ω or, to be more precise, on each event of the "single-element" events $\{HHH\}$, $\{HHT\}$, $\{HTH\}$, etc. For then, the value of $\mathbb{P}$) on larger subsets of Ω follows from repeated use of condition (b) of definition 6.8; for example:

$$\mathbb{P}\left(\{HHT, HTH, HHH\}\right) = \mathbb{P}\left(\{HHT\}\right) + \mathbb{P}\left(\{HTH\}\right) + \mathbb{P}\left(\{HHH\}\right),$$

etc. Now for a single-element event like $\{HTH\}$ (corresponding to throwing first H then T and then H again) it is reasonable to put

$$\mathbb{P}(\{HTH\}) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8},$$

and similarly for the others: $\mathbb{P}(\{HHH\}) = \mathbb{P}(\{HHT\}) = 1/8$, etc. (Note that we're assuming here not only that the coin is fair, bur also that subsequent tosses are independent of each other, that is, the result of the first toss does not influence the second, etc.!) In this way one constructs a probability-measure on $\mathcal{F}_{discr}$.

The previous example is characteristic: in defining probability measures we often first define it on a smaller generating set of the σ-algebra[15], and then extends it to all of the σ-algebra by repeated use of the σ-additivity, property (b) of definition 6.8. For less simple σ-algebras than the discrete one this is somewhat less easy to carry out, but the following abstract result solves this extension problem in a very general setting:

Theorem 6.11. *(Carathéodory's extension theorem) Let $\mathcal{C}$ be an algebra of subsets of Ω, that is, $\mathcal{C}$ is closed under finite unions and taking complements. If*

$$\mu : \mathcal{C} \to [0, \infty],$$

is a map satisfying the following version of σ-additivity:

$$\text{For all } F_1, F_2, \cdots \in \mathcal{C}, \text{ if } \cup_j F_j \in \mathcal{C}, \text{ then } \ \mu\left(\cup_j F_j\right) = \textstyle\sum_j \mu(F_j)\ .$$

Then μ can be extended to a σ-additive map

$$\mu : \mathcal{F}(\mathcal{C}) \to [0, \infty],$$

that is, a map which satisfies condition (b) of definition 6.8.

The proof of this theorem is outside the scope of these lectures, and I am only mentioning it because, in the examples below, we will define probability measures by specifying their values only on specific events which together, however, generate the whole σ-algebra. Carathéodory's theorem can be used to provide a theoretical underpinning for this procedure, of which you won't need to know the details.

Example 6.12. *(Lebesgue measure on $\mathbb{R}$ and $\mathbb{R}^n$)* Let us take $\Omega = \mathbb{R}$, and $\mathcal{F} = \mathcal{B}(\mathbb{R})$, the Borel σ-algebra. A moments thought shows that $\mathcal{B}(\mathbb{R})$ is already generated by the set of al (left-) half open intervals:

$$\{(a, b] : a, b \in \mathbb{R} \text{ or } a, b = \pm\infty \text{ and } a < b\}.$$

(Note that we include infinite intervals like $(-\infty, b]$ and (a, ∞).) To make this into an algebra we define $\mathcal{C}$ as the set of all finite unions:

$$F = (a_1, b_1] \cup \cdots \cup (a_n, b_n] : -\infty \leq a_1 < b_1 \leq a_2 < b_2 \leq \cdots \leq b_n\};$$

[15] it is instructive exercise to convince yourself that $\mathcal{F}_{discr}(\Omega)$ is generated by $\{\{\omega\} : \omega \in \Omega\}$

that is, we simply add all possible finite unions of disjoint intervals. If we define μ on $\mathcal{C}$ by:

$$\mu(F) := \text{length of } F \ = \sum_{j=1}^{n}(b_j - a_j),$$

for an F as above, then one can show, using Carathéodory's theorem, that μ extends to a measure on $\mathcal{B}(\mathbb{R})$, which is called the *Lebesgue measure*. One usually simply writes dx instead of μ.

For unions of intervals their Lebesgue measure is just their total length, in ordinary sense, but we can now also talk about the length of any Borel subset of $\mathbb{R}$ (which can be much more bizarre than a simple interval). This measure is not a probability measure , since the measure of $\Omega = \mathbb{R}$ is not even finite! One can get a probability measure om $\mathbb{R}$ by taking a positive function $f = f(x) \geq 0$ of total intgral 1 on $\mathbb{R}$ (that is, by taking a probability density!) and defining,

$$\mathbb{P}((a,b]) = \int_a^b f(x)dx,$$

extended to finite unions of disjoint intervals $(a_j, b_j]$ by taking the sum of the respective integrals. This extends again to a measure on $\mathcal{B}(\mathbb{R})$, which now is a probability measure, since:

$$\mathbb{P}(\mathbb{R}) = \int_{-\infty}^{\infty} f(x)dx = 1.$$

Another way to get a probability measure out of Lebesgue measure is by restricting Ω to be the interval $[0,1]$, since this has length 1.

The construction of Lebesgue measure carries over to n-dimensional Euclidian space, $\mathbb{R}^n$: we now take as basic building blocks for our σ-algebra unions of a finite number of disjoint n-dimensional cubes:

$$Q = \{x \in \mathbb{R}^n : a_1 < x_1 \leq b_1, \cdots, a_n < x_n \leq b_n\},$$

and define μ on such a cube as simple being its (n-dimensional) volume:

$$\mu(Q) = (b_1 - a_1) \cdot \cdots \cdot (b_n - a_n).$$

The resulting measure on $\mathcal{B}(\mathbb{R}^n)$ is called, unsurprisingly, n-dimensional Lebesgue measure, and denoted by $dx = dx_1 \cdots dx_n$.

Example 6.13. As a more singular example of a probability measure we again take $\Omega = \mathbb{R}$ and $\mathcal{F} = \mathcal{B}(\mathbb{R})$. We fix a point $x_0 \in \mathbb{R}$, and define a measure δ_{x_0} by:

$$\delta_{x_0}(B) = \begin{cases} 1 \text{ if } x_0 \in B, \\ 0 \text{ otherwise} \end{cases}$$

This is called the *Dirac delta-measure in* x_0, and models a situation where we are sure that the "state of the world" is x_0.

As a generalization, we can take a sequence of points $x_j \in \mathbb{R}$, and a sequence of positive numbers $p_j \geq 0$, and define

$$\mathbb{P}(B) = \sum_{j:x_j \in B} p_j.$$

If $\sum_{j=0}^{\infty} p_j = 1$, this defines a probability measure on $\Omega = \mathbb{R}, \mathcal{F} = \mathcal{B}(\mathbb{R})$. A concrete example is given by:

$$x_j = j, \quad p_j = \frac{\lambda^j}{j!} e^{-\lambda},$$

and for example $\mathbb{P}((a, b]$ is then simple the probability that a Poisson random variable will have its value in $(a, b]$.

Definition 6.14. A triple $(\Omega, \mathcal{F}, \mathbb{P})$ with $\mathcal{F}$ a σ-algebra of subsets of Ω and $\mathbb{P}$ a probability measure on Ω will be called a *probability space.*

6.4. **Random variables.** A real random variable can be thought of as an object which can take on different values in different (future) states ω of the world, so we can simply look upon it as being a function

$$X : \Omega \to \mathbb{R}$$

or *define* it as such. However, it cannot be any function; we would like to be able to consider events like: "X will take on a value between a and b", or $a < X < b$, and talk about its probability:

$$\mathbb{P}(a < X < b).$$

The set of "possible future states of the world" in which X lies between a and b is given by:

$$\{\omega \in \Omega : a < X(\omega) < b\}.$$

This is simply the *inverse image of* $(a, b <$ *under* X, and often denoted by: $X^{-1}((a, b <)$. More generally, for any set $G \subset \mathbb{R}$, we put[16]

$$X^{-1}(G) = \{\omega \in \Omega : X(\omega) \in G\};$$

To be able to assign probability to them, these sets should lie in $\mathcal{F}$, and we therefore arrive at the following important definition:

Definition 6.15. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A *real random variable* is a function $X : \Omega \to \mathbb{R}$ such that:

$$X^{-1}((a, b)) \in \mathcal{F}, \text{ for all } a, b \in \mathbb{R}. \tag{138}$$

Functions which satisfy (138) are also called *measurable w.r.t.* $\mathcal{F}$, or $\mathcal{F}$*-measurable.* We note in passing that one can show that if X is measurable, then in fact the inverse image $X^{-1}(B)$ of any Borel subset of $\mathbb{R}$ will be in $\mathcal{F}$. In particular, we can take $B = (a, b]$, where $a = -\infty$ is allowed: see exercise 6.22.

[16] As a perhaps totally superfluous remark, we emphasize that this should *not* be confused with the notation X^{-1} for the *inverse function* of X, which may or may not exist, and which anyhow, in the present situation, only makes sense if $\Omega = \mathbb{R}$!

The following definition now makes the connection between the present formalism and the informal approach to probability of the first chapters:

Definition 6.16. If X is a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then its *cumulative distribution function* is the function $F_X : \mathbb{R} \to [0,1]$ defined by:

$$F_X(x) = \mathbb{P}\left(X^{-1}((-\infty, x]) \right) = \mathbb{P}\left(\{\omega \in \Omega : X(\omega) \leq x\}\right). \tag{139}$$

Indeed, the right hand side of (139) is simply the probability of the event that $X \leq x$, written down in our new formalism.

Vector-valued random variables are defined similarly:

Definition 6.17. A *vector-valued random variable on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ is a function $\mathbb{X} \to \mathbb{R}^n$, for some given n, such that, for all $a_1, b_1, \cdots, a_n, b_n$,

$$\begin{aligned} &\mathbb{X}^{-1}\left((a_1, b_1) \times \cdots \times (a_n, b_n)\right) \\ &(= \{\omega : a_1 < X_1(\omega) < b_1, \cdots, a_n < X_n(\omega) < b_n\}) \\ &\in \mathcal{F}, \end{aligned}$$

where the X_j are the components of $\mathbb{X} = (X_1, \cdots, X_n)$.

Again, one can show that if $\mathbb{X}$ is a vector random variable, then $\mathbb{X}^{-1}(B) \in \mathcal{F}$, for each Borel set $B \in \mathcal{B}(\mathbb{R}^n)$. The cumulative distribution-function of $\mathbb{X}$ is now defined as

$$F_{\mathbb{X}}(x) = \mathbb{P}\left(\{\omega : X_1(\omega) < x_1, \cdots, X_n(\omega) < x_n\}\right), \tag{140}$$

if $x = (x_1, \cdots, x_n) \in \mathbb{R}^n$.

6.5. **Brownian motion revisited.** We illustrate the formalism introduced above by re-examining Brownian motion from our new measure-theoretic point of view.

1. The sample space. We take as our $\Omega_{Brownian}$ the set of all real-valued functions on the positive half-line:

$$\Omega_{Brownian} = \{w : [0, \infty) \to \mathbb{R}\},$$

2. The σ-algebra. We define $\mathcal{F}_{Brownian}$ as the σ-algebra generated by all sets of the form:

$$F = F_{(t_j, a_j, b_j)_{1 \leq j \leq N}} := \{w : [0, \infty) \to \mathbb{R} : w(t_1) \in (a_1, b_1), , \cdots, w_N \in (a_N, b_N)x\}, \tag{141}$$

where N runs over $\mathbb{N}$, $0 \leq t_1 < t_2 < \cdots < t_N$ and a_j and b_j are arbitrary elements of $\mathbb{R}$. This σ-algebra has a similar explicit description as the

one in example 6.6, where now we only have to restrict the Borel-sets in (135) to subsets of the positive reals: $B_j \subset [0, \infty)$.

3. The random variables. Brownian motion will now be defined as the collection of random variables W_t , $t \geq 0$, defined as:

(142) $W_t(\omega) = w(t)$ if $\omega \in \Omega_{Brownian}$ is the function $w : [0, \infty) \to \mathbb{R}$.

4. The probability measure. Finally, for the definition of our Brownian probability measure $\mathbb{P}_{Brownian}$ we take our inspiration from (50), and put for F defined by (141):

$$\mathbb{P}_{Brownian}(F) = \int_{a_1}^{b_1} dx_1 \cdots \int_{a_N}^{b_N} dx_N \tag{143}$$
$$p_0(x_N, t_N - t_{N-1}) p_0(x_{N-1} - x_{N-2}, t_{N-1} - t_{N-2}) \cdots p_0(x_1, t_1),$$

where, as before,

$$p_0(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}.$$

The motivation for this complicated looking definition of $\mathbb{P}_{Brownian}$ is that, if W_t is defined by (142), then F given by (141) is precisely the event that

$$W_{t_1} \in (a_1, b_1] \text{ and } W_{t_2} \in (a_2, b_2] \text{and } \cdots \text{ and} W_{t_N} \in (a_N, b_N],$$

whose probability should be given by integrating the joint probability density (50).

Carathéodory's theorem now allows us to extend $\mathbb{P}_{Brownian}$ to the whole of our σ-algebra $\mathcal{F}_{Brownian}$.

One can now easily convince oneself that the Brownian motion $(W_t)_t$ defined on $\Omega_{Brownian}, \mathcal{F}_{Brownian}, \mathbb{P}_{Brownian}$ satisfies the defining properties (i), (ii) and (iii) of Brownian motion.

Formula's like (143) admittedly look rather unappetizing, but we will in fact very seldom work directly on the probability space ($\Omega_{Brownian}$, $\mathcal{F}_{Brownian}, \mathbb{P}_{Brownian}$) (or any other probability space, for that matter) when using Brownian motion in practical applications. The new measure-theoretic view of probability should be looked upon as a convenient theoretic framework. One of its main conceptual advantages is the separation of the set of events $\mathcal{F}$ and of the probability $\mathbb{P} : \mathcal{F} \to [0, 1]$ given to those events. Indeed, one easily imagines two investors looking at the same set of events involving a stock's price, but assigning different probabilities to them. Changing probabilities will play an important rôle in Pricing II, where we will see that derivative securities are priced as the (discounted) expectation of their pay-offs, not under the "real-world" probability, but under a risk-neutral one.

*__Remark 6.18.__ *(To be ignored unless you know a bit of Analysis)* What about property (iv), continuity? This turns out to be considerably more subtle. We would like to show that the event of being a continuous path has probability 1. In mathematical language, if

$$F_{CP} = \{w : [0,\infty) \to \mathbb{R} : \ w \text{ continuous at all its points}\},$$

("CP" standing for "Continuous Pats") then $\mathbb{P}_{Brownian}(F_{CP}) = 1$. However, a first problem is that the event F_{CP} is not even in our σ-algebra $\mathcal{F}_{Brownian}$! See the (non-mandatory) exercises at the end of this chapter for an explication: the point is that continuity at all $t \geq 0$ imposes an uncountable set of conditions. The way out is to first show that the event $F_{UC\mathbb{Q}}$ of being *uniformly continuous on* $\mathbb{Q}_{\geq 0} \cap [0,T]$*, each* $T > 0$*, is* in the σ-algebra, and has total probability 1 with respect to $\mathbb{P}_{Brownian}$ (the latter is not at all easy to show!). We then re-define our Brownian motion by:

$$W_t^*(\omega) = \lim_{\substack{r \to t \\ 0 \leq r \in \mathbb{Q}}} W_r(\omega), \text{ if } \omega \in F_{UC\mathbb{Q}},$$

(the limit exists because of uniform continuity) while we simply put

$$W_t^*(\omega) = 0, \text{ if } \omega \notin F_{UC\mathbb{Q}}.$$

All sample paths $t \to W_t^*(\omega)$ will now be continuous. Moreover, W_t^* will have the same properties (i), (ii) and (iii) as W_t, and is therefore or sought-for Brownian motion wit continuous sample paths. At this point we simply write again W_t for W_t^*. One can show that, for each fixed $t \geq 0$, $W_t(\omega) = W_t^*(\omega)$, with probability 1. All this theory is explained in detail in the more mathematically oriented books on probability theory, like for example:

P. Billingsly, Probability and Mesure, Wiley-Interscience Publication, John Wiley & Sons, 3-th ed., 1995,

N. V. Krylov, Introduction to the Theory of Random Processes, Graduate Studies in Mathematics, Volume 43, American Matematical Society, 2002.

L. C. G. Rogers and D. Williams, Diffusions, Markov Processes and Martingales, Wiley Series in Probability and Math. statistics, 2-nd ed., 1994

6.6. **σ-algebras generated by random variables.** If $X : \Omega \to \mathbb{R}$ is a random variable, we let:

$$\sigma(X) := \left(\begin{array}{c} \text{smallest } \sigma\text{-algebra containing all} \\ \text{sets } X^{-1}((a,b)),\ a,b \in \mathbb{R} \end{array} \right), \tag{144}$$

the *sigma-algebra generated by* X. The idea is, that σ_X contains all possible information about which state of the world we are in which

can be obtained from observation of the random variable X. One can show that:

$$\begin{aligned} \sigma(X) &= X^{-1}\left(\mathcal{B}(\mathbb{R})\right) \\ &:= \{X^{-1}(B) : B \in \mathcal{F}\}. \end{aligned} \tag{145}$$

We will need to go beyond one rv, and consider the σ-algebra generated by an entire families of random variables. The most important example of such a family is that stochastic process, $(X)_t, t \geq 0$. We then let:

$$\sigma(X_s : s \leq t) = \left(\begin{array}{c} \text{smallest } \sigma\text{-algebra containing all sets} \\ X_s^{-1}((a,b)), \text{ for } s \leq t \text{ and } a, b \in \mathbb{R} \end{array} \right), \tag{146}$$

interpreted as *the information about the state of the world, contained in the process X up till, and including, time t.*

Example 6.19. *(Brownian motion again)* We take up the example of Brownian motion again, realized on the probability space $(\Omega_{Brownian}, \mathcal{F}_{Brownian}, \mathbb{P}_{Brownian})$ introduced in the previous subsection. If we *fix* a time t, then one easily guesses that $\sigma(W_t)$ is the σ-algebra consisting of all sets of the form

$$\sigma(W_t) = \{w : [0,\infty) \to \mathbb{R} : w(t) \in B\},$$

where B ranges over the Borel subsets of $\mathbb{R}$. This can be sown formally using (145) and the definition of W_{t_0} on $\Omega_{Brownian}$.

More generally, $\sigma(W_s, s \leq t)$ can, in this realization of Brownian motion, be given a more concrete description as the collection of all sets F of the form

$$F = \{s : [0,\infty) \to [0,\infty) : s(t_j) \in B_j,\ j = 1, 2, \cdots\},$$

where $s_1 \leq s_2 \leq \cdots \leq t$ is a sequence of times *smaller or equal to t*, and where B_j is a Borel subset of $\mathbb{R}$.

As already noted, such explicit descriptions are useful for illustrative purposes only, to give an idea of what the general definition means in a particular case. In practice, it is much easier to work with some abstract realization of Brownian motion $(W_t)_{t\geq 0}$ on some unspecified probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The σ-algebra generated by Brownian motion up till time t will often be denoted by $\mathcal{F}_t$, or $\mathcal{F}_t^W$, if we want to stress Brownian motion:

$$\mathcal{F}_t = \mathcal{F}_t^W = \sigma\left(W_s : s \leq t\right). \tag{147}$$

These σ-algebras will play an extremely important rôle in the construction of stochastic integrals: basically, we will be able to integrate

$$\int_0^T H_t dW_t,$$

with *stochastic* integrands H_t, that is, the H_t are rv $H_t : \Omega \to \mathbb{R}$, but *only those such that, for each* t*,*

$$H_t : \Omega \to \mathbb{R},$$

is $\mathcal{F}_t^W$*-measurable.* In fact, there is a slight subtlety with the definition of $\mathcal{F}_t$, in that we agree to also include all events of probability 0, that is, events which are never expected to produce themselves with Brownian motion.

6.7. **Events of probability 0 or null-events.** We say that $F \in \mathbb{F}$ is a *null-event* or *null-set with respect to* $\mathbb{P}$ if

$$\mathbb{P}(F) = 0.$$

This does not mean that F is impossible, but simply that it practically will not occur.

It is not important here that $\mathbb{P}$ is normalized ($\mathbb{P}(\Omega) = 1$), and we can define in the same way a null-set with respect to a measure.

Examples 6.20. (i) If we take $\Omega = [0, 1]$, with the Borel σ-algebra (the one generated by the intervals), and with Lebesgue-measure, then single-element sets $\{x_0\}$ are null-events: indeed, $\{x_0\} = [x_0, x_0]$ and $\mathbb{P}([x_0, x_0]) = x_0 - x_0 = 0$. More generally, an infinite but discrete set $\{x_0, x_1, x_2, \cdots\}$ is a null event: by the σ-additivity of $\mathbb{P}$,

$$\mathbb{P}\left(\{x_0, x_1, x_2, \cdots\}\right) = \sum_j \mathbb{P}(\{x_j\}) = 0.$$

We remark in passing that null-events in $[0, 1]$ can be a lot bigger than just a discrete sequence.

(ii) In the context of the Brownian motion example **??**, of $t_0 > 0$ and $x_0 \in \mathbb{R}$,

$$\{\omega \in \Omega_{Brownian} : W_{t_0} = x_0\} =: (W_{t_0} = x_0)$$

is a null-event (exercise!). More generally, $(W_{t_0} \in F_0)$ is a null-event for $\mathbb{P}_{Brownian}$ if $F_0 \subset \mathbb{R}$ is a null-set for Lebesgue measure.

One often extends the concept of null-event to arbitrary subsets of Ω, by saying that a subset $A \subset \Omega$ is a null-event or null-set, if there exists an $F \in \mathbb{F}$ such that:

$$A \subset F \text{ and } \mathbb{P}(F) = 0.$$

The slightly subtle point here is that, a priori, A itself need not be an element of $\mathcal{F}$, so that we cannot directly speak of $\mathbb{P}(A)$. One then usually enlarges $\mathcal{F}$ by throwing in all the null-sets defined in this way, and completing to a σ-algebra. In the classical context of the Borel σ-algebra, one obtains in this way a new σ-algebra which is called the σ-algebra of *Lebesgue-measurable sets.*

By convention, we do something similar with the $\mathcal{F}_t^W$, by adding all null-sets.

6.8. **Appendix to chapter 6: brief review of set-theoretic notations.** We need to work with sets of various mathematical objects, like sets of real numbers, sets of vectors in $\mathbb{R}^n$, but also more complicated ones, like sets of subsets of real numbers (σ-algebras!), or the sets of functions from the positive reals to the reals.

Elements of a set. The members of a set are called its *elements*, and we use

$$x \in A,$$

for "x belongs to A", or "x is an element of A".

Subsets. If A and B are sets, then

$$A \subseteq B,$$

means that every element of A is an element of B, and we say that A *is a subset of* B.

Functions and inverse images. The notation

$$f : A \to B$$

means that f is a *function* from a set A to a set B, that is, an operation which maps each element of A to a single element of B. An important notion for us will be that of the *inverse image*,

$$f^{-1}(C),$$

of a subset $C \subseteq B$ of B:

$$f^{-1}(C) = \{a \in A : f(a) \in C\},$$

the set of all elements of a which are mapped , by f, to an element of C.

Inverse functions. The notation for inverse image should not be confounded with that for an *inverse function*, f^{-1}: if $f : A \to B$ is *one-to-one and onto*[17], that is, if for all $a_1, a_2 \in A$,

- $f(a_1) = f(a_2) \Rightarrow a_1 = a_2$
  (f one-to-one or *injective* ,
- Each element $b \in B$ is of the form $f(a)$, for some $a \in A$
  (f onto, or *surjective*),

then we can define the inverse map $f^{-1} : B \to A$ by:

$$f^{-1}(b) = a \Leftrightarrow f(a) = b.$$

We won't very much use inverse functions, but inverse images will occur a lot in what we will do.

Operations on sets: Given two sets A and B, we can form their *intersection*, their *union*, and their *difference*. These are defines as follows:

[17] such functions are often called *bijective* in the mathematics literature

Intersection: $A \cap B = \{x : x \in A \text{ and } x \in B\}$, the set of all elements which are both in A and in B.

Union: $A \cup B = \{x \in A \text{ or } x \in B\}$, the set of all elements which are in both A and B.

Difference: $A \setminus B = \{x \in A : x \notin B\}$, the set of elements of A which are not in B. Note that this will in general be different from $B \setminus A$!

Notations for some standard subsets of $\mathbb{R}$ (the set of all real numbers):

$$(a,b) = \{x \in \mathbb{R} : a < x < b\} \quad \text{(open interval)},$$

$$[a,b] = \{x \in \mathbb{R} : a \leq x \leq b\} \quad \text{(closed interval)}$$

$$(a,b] = \{x \in \mathbb{R} : a < x \leq b\} \quad \text{(half-open interval to the left)},$$

$$[a,b) = \{x \in \mathbb{R} : a \leq x < b\} \quad \text{(half-open interval to the right)}$$

We also put:

$$(a,\infty) = \{x \in \mathbb{R} : x > a\},$$

and

$$[a,\infty) = \{x \in \mathbb{R} : x \geq a\};$$

$(-\infty, b)$ and $(-\infty, b]$ are defined similarly. For example,

$$(-\infty, b) = \{x \in \mathbb{R} : x < b\},$$

etc.

6.9. *__Exercises for chapter 6.__ The following exercises are somewhat more theoretical than you may be used to, or indeed than what is in general required for Financial Engineering practice. They have only been included to illustrate certain ways of reasoning with σ-algebras, and to justify some of the claims made in the text, and they do not constitute examinable material.

Exercise 6.21. (a) Show that if $F_1, F_2, \cdots, F_n, \cdots$ is a finite or infinite collection of subsets of Ω, then

$$\cup_n \left(\Omega \setminus F_n\right) = \Omega \setminus \left(\cap_n F_n\right).$$

(b) Use a) to show that if $\mathcal{F}$ is a σ-algebra and $F_1, F_2. \cdots \in \mathcal{F}$, then also $\cap_n F_n \in \mathcal{F}$.

Exercise 6.22. a) Let $a \leq b$. Show that $[a,b]$, $(a,b]$ and $[a,b)$ are all in $\mathcal{B}(\mathbb{R})$.
Hint: show that, for example, $(a,b] = \cap_{n=0,1,\cdots}(a, b + \frac{1}{n})$, and similarly for the others.)

b) Show that $(-\infty, a)$ and (b,∞) are in $\mathcal{B}(\mathbb{R})$.

c) Show that if X is a real random variable on $(\Omega, \mathcal{F}, \mathbb{P})$, then $X^{-1}((a,b]) \in \mathcal{F}$.
(*Hint:* verify first that, very generally, $X^{-1}(\cap_n B_n) = \cap_n X^{-1}(B_n)$.)

Exercise 6.23. Show that if X is a random-variable on X, then $X^{-1}(B) \in \mathcal{F}$, for all $B \in \mathcal{B}(\mathbb{R})$, by completing the following steps:

a) Let $\mathcal{G} = \{G \subset \mathbb{R} : X^{-1}(G) \in \mathcal{F}\}$. Show that $\mathcal{G}$ is a σ-algebra on $\mathbb{R}$.

b) Explain why $(a, b) \in \mathcal{G}$, for all $a, b \in \mathbb{R}$.

c) Use a), b) and the definition of $\mathcal{B}(\mathbb{R})$, to conclude that $\mathcal{B}(\mathbb{R}) \subset \mathcal{G}$. From this conclude the desired property of X with respect to the Borel sets.

Exercise 6.24. Prove the affirmation in example 6.5 that the σ-algebra generated by the sets (133) is precisely the set:

$$\mathcal{F}_B = \{B_1 \times B_2 \times \cdots : B_j \in \mathcal{B}(\mathbb{R}_{\geq 0}\}. \tag{148}$$

(The subscript "B " stands for "Borel").
(*Hint:* To prove such a result, it suffices to prove that:

- $\mathcal{F}_B$ is itself already a σ-algebra ,
- Any σ-algebra which contains the sets (133) will necessarily contain $\mathcal{F}_B$.

For the latter point, first convince yourself that any σ-algebra which contains the sets (133) must contain all events of the form

$$F_{j,B_j} = \{(s_0, s_1, s_2, \cdots) : s_j \in B_j\}$$

$$= \mathbb{R}_{\geq 0} \times \cdots \times B_j \times \cdots \times \mathbb{R}_{\geq 0} \times \cdots ,$$

that is, the set of points $\omega = (s_0, s_1, s_2, \cdots)$ in sample space whose j-th coordinate is in the set $B \in \mathcal{B}(\mathbb{R}_{\geq 0})$.

Exercise 6.25. Similarly, in the context of example 6.6, prove the affirmation in example 6.6, that the σ-algebra generated by the sets (134) is precisely of the set of all events of the form (135) .

Exercise 6.26. Prove (145 by showing that:

a) $X^{-1}(\mathcal{B}(\mathbb{R})$ is already a σ-algebra in its own right.

b) By an argument similar to the one used in exercise 6.23, show that any sub-σ-algebra $\mathcal{F}_1$ of $\mathcal{F}$ containing all $X^{-1}((a, b))$ has to contain all of $X^{-1}(\mathcal{B}(\mathbb{R}))$. (It suffices in fact to replace $\mathcal{F}$ by $\mathcal{F}_1$ in that exercise.)

c) Using the formal definition of $\sigma(X)$ as the common intersection of all sub-sigma algebras $\mathcal{F}_1$ containing all $X^{-1}(a, b)$, conclude.

Exercise 6.27. *On a mathematical difficulty with being continuous with probability 1.*

(This exercise supposes you are familiar with basic concepts form Mathematical Analysis, and should simply be skipped if you're not).

We take the sample-space $\Omega = \Omega_{Brownian}$ and the σ-algebra $\mathcal{F} = \mathcal{F}_{Brownian}$. Points of Ω are thus functions $w : [0, \infty) \to \mathbb{R}$, and events are certain sets of such functions.

Recall that a function $w : [0, \infty) \to \mathbb{R}$ is called *continuous in the point* t_0 if, for all $\varepsilon > 0$ there exists a $\delta > 0$ such that, for all $t \in [0, \infty)$:

$$|t - t_0| < \delta \Rightarrow |w(t) - w(t_0)| < \varepsilon.$$

We can, without loss of generality, limit ourselves to ε and δ of the form $1/n$, $1/m$, $n, m \in \mathbb{N}$. Continuity at t_0 can therefore be expressed as:

$$\text{For all } n \text{ there exists an } m \text{ such that } |t - t_0| < 1/m$$
$$\Rightarrow |s(t) - s(t_0)| < 1/n.$$

(a) Show that the event "$w \in \Omega$ continuous in $t = 1$" corresponds to the following subset of Ω:

$$\cap_{n\in\mathbb{N}} \cup_{m\in\mathbb{N}} \cap_{t\geq 0,\ |t-t_0|<1/m} \{w \in \Omega : |w(t) - w(1)| < 1/n\}.$$

Explain why this set (probably) does not belong to $\mathcal{F}$. Can you actually *prove* this? (see also part (b) below).

(b) We can give a similar description of the σ-algebra $\mathcal{F}$ as for the one of example 6.6 (see previous exercise). In particular, any event $F \in \mathcal{F}$ is of the form:

$$F = \{w \in \Omega : w(t_1) \in B_1, w(t_2) \in B_2, \cdots\},$$

for a countable set of times t_j and Borel sets $B_j \in \mathcal{B}(\mathbb{R})$, and whether a w is in F or not is completely determined by what happens with w at this discrete set of times t_j, $j = 1, 2, \cdots$. Use this observation to argue that the set:

$$\{w : [0, \infty) \to \mathbb{R} : w \text{ continuous at all } t \geq 0\ \},$$

is *not* in $\mathcal{F}$.

The conclusion of (a) and (b) is that events like "being continuous at a point or at all points" are not observable in the σ-algebra $\mathcal{F}$. So how come we can talk about Brownian motion being continuous with probability 1? The key is by first looking at Brownian motion restricted to *rational* positive times:

$$\mathbb{Q}_+ = \{r = p/q : p, q \in \mathbb{N}\}.$$

(c) Show that the set:

$$F_C := \{w : [0, \infty) \to \mathbb{R} : w\big|_{\mathbb{Q}_+} \text{ is continuous in all positive rationals}\}$$

is an element of $\mathcal{F}$, by showing that it equals:

$$\cap_{r\in\mathbb{Q}_+} \cap_{n\in\mathbb{N}} \cup_{m\in\mathbb{N}} \cap_{r'\in\mathbb{Q}_+, |r-r'|<1/m} \{w \in \Omega : |w(r) - w(r')| < 1/n\}.$$

More generally, show that the event $F_{UC\mathbb{Q}}$ from remark 6.18 is in $\mathcal{F}$.

7. Expectations and Integrals.

If $X : \Omega \to \mathbb{R}$ is a random variable in the newly defined sense, we would like to know how to compute its expectation. The strategy is to first define the expectation for a set of rvs having a particular simple form, and then to try extend to a general rv by approximation. We will often use the very convenient concept of the *indicator function* of a subset $F \subset \Omega$. This is the function $\mathbb{I}_F : \Omega \to \mathbb{R}$ defined by:

$$\mathbb{I}_F(\omega) = \begin{cases} 1 \text{ if } \omega \in F, \\ 0 \text{ otherwise} \end{cases}$$

7.1. **Defining expectations.** A random variable $X : \Omega \to \mathbb{R}$ is called *simple* if there exist finitely many mutually exclusive sets $F_1, \cdots, F_k \in \mathbb{F}$ and finitely many real numbers $c_1, \cdots, c_k$ such that:

$$X(\omega) = \sum_{j=1}^{k} c_j \mathbb{I}_{F_j}(\omega). \tag{149}$$

As a random variable, X takes on the value c_j in case of the event F_j; since the events are mutually exclusive, there is no ambiguity. We now define the *integral* of such a simple function by:

$$\int_\Omega X(\omega) d\mathbb{P}(\omega) := \sum_{j=1}^{k} c_k \mathbb{P}(F_k). \tag{150}$$

Observe thta

$$\begin{aligned} \int_\Omega X(\omega) d\mu(\omega) &= \sum_{j=1}^{k} c_k \mathbb{P}(F_k) \\ &= \sum_{j=1}^{k} c_j \cdot (\text{Probability that } X = c_j), \end{aligned}$$

so that the integral of X is the same as the expectation of X:

$$\int_\Omega X(\omega) d\mathbb{P}(\omega) = \mathbb{E}(X).$$

So expectations can be regarded as integrals over sample space, with respect to the probability measure: this is conceptually a very powerful point of view, the more so since integrals can also be defined with respect to measures μ which are not probability measures.

We observe in passing that $X = \mathbb{I}_F$ is a special example of a simple function, and that expectation, integral and probability all coincide for such X:

$$\mathbb{E}(\mathbb{I}_F) = \int_\Omega \mathbb{I}_F(\omega) d\mathbb{P}(\omega) = \mathbb{P}(F).$$

We still have to go beyond integration of simple functions. The idea is to do this by *approximation*: if $X : \Omega \to \mathbb{R}$ is an arbitrary $\mathcal{F}$-measurable *positive* function, then one can find increasing sequences of simple functions Y_n $(n = 1, 2, 3, \cdots)$ such that, for any $\omega \in \Omega$,

$$\begin{aligned} &Y_n(\omega) \leq Y_{n+1}(\omega), \quad n = 1, 2, \cdots ; \\ &Y_n(\omega) \to Y(\omega) \text{ as } n \to \infty, \end{aligned} \tag{151}$$

and one puts

$$\int_\Omega Y d\mathbb{P} = \lim_{n\to\infty} \int_\Omega Y_n(\omega) d\mathbb{P}(\omega), \tag{152}$$

where the limit in the right hand side exists (though it might be $+\infty$), and can be shown to be independent of the choice of approximating sequence f_n (this is not trivial!). One can also show that in the case of $\Omega = [0, 1]$, and a continuous function $X = f : [0, 1] \to \mathbb{R}$, and $\mathbb{P} = dx$, then one obtains the familiar Riemann-integral from elementary calculus:

$$\int_0^1 f(x)dx,$$

which can be computed by finding a primitive, etc.

For X's which are not positive, one writes $X(\omega) = X_+(\omega) - X_-(\omega)$of its positive and negative parts[18],

$$X_+(\omega) = \max(X(\omega), 0), \quad X_-(\omega) = \max(-X(\omega), 0),$$

and puts:

$$\begin{aligned} \mathbb{E}(X) &= \int_\Omega X d\mathbb{P} = \int_\Omega X_+ d\mu - \int_\Omega X_- d\mathbb{P} \\ &= \mathbb{E}(X_+) - \mathbb{E}(X_-) \end{aligned}$$

provided both of the terms on the right are finite (to circumvent problems with expressions like $\infty - \infty$, which we can't give a meaning to).

We define the integral of the rv X over smaller sets $F \in \mathcal{F}$, by simply multiplying the integrand by the indicator function of F; that is, we put:

$$\begin{aligned} \int_F X \cdot \mathbb{I}_F \, d\mathbb{P} &:= \int_\Omega X \cdot \mathbb{I}_F \, d\mathbb{P} \\ &= \int_\Omega X(\omega) \cdot \mathbb{I}_F(\omega) \, d\mathbb{P}(\omega). \end{aligned} \tag{153}$$

We can also write (153) as

$$\mathbb{E}\left(X \cdot \mathbb{I}_F\right),$$

[18] if $X : [0, 1] \to \mathbb{R}$, then X_+ and $-X_-$ are simple the part of the graph of X above and below the x-axis, respectively

the expectation of X restricted to the event F.

The following fact is useful to know, and intuitively obvious:

$$F \text{ a null-event } \Rightarrow \int_F X \, d\mathbb{P} = 0. \tag{154}$$

7.2. ***Connection with the older definition.** In the remainder of this section we will explain why this new definition of expectation amounts to the same thing as the previous one involving distribution functions, which we gave in chapter 1. This subsection may be skipped, and the result simply admitted.

It suffices to treat the case of positive rvs X: the general case can be handled by writing X as $X_+ - X_-$, as we did above. Pick an $n \in \mathbb{N}$, to be thought of as a very big number (it will tend to ∞ in te end), and divide the real line $\mathbb{R}$ in small intervals of size 2^{-n}:

$$(\frac{j}{2^n}, \frac{j+1}{2^n}], \quad j \in \mathbb{Z} = \{\cdots, -2, -1, 0, 1, 2, \cdots\}.$$

Put

$$X_n(\omega) = \sum_j \frac{j}{2^n} \mathbb{I}_{X^{-1}((j2^{-n},(j+1)2^{-n}])}.$$

(Make a drawing with the graph of X when for example $\Omega = [0, 1]$!) Then one can check (exercise, or see the literature) that

- $X_n(\omega) \leq X_{n+1}(\omega)$;
- $X_n(\omega) \to X(\omega)$, as $n \to \infty$.

It follows that:

$$\begin{aligned} \mathbb{E}(X) &= \lim_{n\to\infty} \int_\Omega X_n(\omega) d\mathbb{P}(\omega) \\ &= \lim_{n\to\infty} \sum_j \frac{j}{2^{-n}} \mathbb{P}\left(\left\{\omega : \frac{j}{2^{-n}} < X(\omega) \leq \frac{j+1}{2^{-n}}\right\}\right) \end{aligned} \tag{155}$$

Now

$$\left\{\omega : \frac{j}{2^{-n}} < X(\omega) \leq \frac{j+1}{2^{-n}}\right\} = \left\{\omega : X(\omega) \leq \frac{j+1}{2^{-n}}\right\} \setminus \left\{\omega :\leq \frac{j}{2^{-n}}\right\}, \tag{156}$$

and if we apply the general rule that, for any $F_1, F_2 \in \mathcal{F}$ such that $F_2 \subset F_1$,

$$\mathbb{P}(F_1 \setminus F_2) = \mathbb{P}(F_1) - \mathbb{P}(F_2),$$

and recall the definition of the distribution function, (139), we see that the probability of (156) is simply

$$F_X\left(\frac{j+1}{2^n}\right) - F_X\left(\frac{j}{2^n}\right),$$

and that therefore (155) implies that:

$$\mathbb{E}(X) = \lim_{n\to\infty} \sum_j \frac{j}{2^n}\left(F_X\left(\frac{j+1}{2^n}\right) - F_X\left(\frac{j}{2^n}\right)\right), \tag{157}$$

which in chapter 1 we denoted by:

$$\int_{\mathbb{R}} x\ dF_X(x);$$

see formulas (11), (12) with $g(x) = x$.. We can do the case of an arbitrary (positive and increasing) $g = g(x)$ in a similar way, by observing that $g(X)$ is the limit of the following sequence of simple functions:

$$\sum_j g\left(\frac{j}{2^n}\right)\mathbb{I}_{X^{-1}((j2^{-n},(j+1)2^{-n}])},$$

and repeating the argument. More general g then are handled by writing them as difference of increasing functions (as an exercise you might like to think about why this is always possible if g is (continuously) differentiable).

These expressions remain of course rather theoretical, and they will only become more concrete if we suppose for example that F_X is differentiable, $F_X' = f_X$, in which case $dF_X = f_X dx$. Observe, however, that if we would be doing some kind of non-parametric statistics, and our only knowledge of X would be in the form a histogram, then we would basically know the $F_X(j/2^n)$ for some fixed n, and formulas like (157) and its generalization for $g(X)$, (11) are used (without the limit) to compute approximate values of the respective means. This remark could apply for example to the pricing of over the counter options when one has done a non-parametrical regression of the relevant risk-neutral distribution on the basis of the prices of liquidly traded vanilla options.

The above generalizes to multi-variate distribution functions, and the results agree with what we did in chapter 1; we won't go into details.

8. **Hilbert Spaces**

It is useful, both for stochastic calculus, as for Finance in general, to know the basics of the theory of Hilbert spaces. Indeed, the set of random variables of finite variance on a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a prime example of a Hilbert space. In Analysis, this space is known as the space of square integrable functions, $L^2(\Omega, \mathcal{F}, \mathbb{P})$. Hilbert space terminology permeates all of modern mathematics, from the theory of differential equations (where it originated) to probability theory, statistics and mathematical finance.

8.1. **pre-Hilbert and Hilbert spaces.** A Hilbert space is basically an infinite-dimensional generalization of the familiar Euclidean spaces $\mathbb{R}^n$ (we will only be concerned with what are called *real* Hilbert spaces). We give a slightly informal and descriptive, rather than a formal definition, for which we refer to the mathematical literature.

Definition 8.1. A *pre-Hilbert space* H is a real vector space, provided with an *inner product*, which is a map from $H \times H$ to $\mathbb{R}$, sending pairs of elements x, y to a number $(x, y)_H \in \mathbb{R}$, such that for $x, y, z \in H$ and $\lambda \in \mathbb{R}$,

$$(x+y, z)H_=(x, z)_H+(y, z)_H,\ (\lambda\cdot x, y)_H = \lambda(x, y)_H(\lambda, (x, y)_H = (y, x)_H,$$

and

$$(x, x) \geq 0,\ (x, x) = 0 \Leftrightarrow x = 0.$$

That H is a vector space means that we can add elements x, y of H to $x + y \in H$, and multiply them by real numbers λ: $\lambda \cdot x \in H$, in such a way that all the usual rules of algebra are satisfied. We usually simply write te inner product as (x, y) instead of $(x, y)_H$, when there is no confusion possible.

Examples 8.2. Examples of pre-Hilbert spaces are:

(i) $\mathbb{R}^n$, with inner product $(x, y) = \sum_{j=1}^n x_i y_i$.

(ii) The space of square integrable random functions:

$$L^2(\Omega, \mathcal{F}, \mathbb{P}) \tag{158}$$
$$= \{X : \Omega \to \mathbb{R} : X\ \mathcal{F} - \text{measurable}, \int_\Omega X(\omega)^2 d\mathbb{P}(\omega) < \infty\},$$

which is the same as the vector space of random variables X such that $\mathbb{E}(X^2) < \infty$. The inner product is:

$$(X, Y)_{L^2} = \int_\Omega X(\omega)Y(\omega)d\mathbb{P}(\omega), \tag{159}$$

which can be written in a more simple, and also more probabilistic, way as:

$$(X, Y)_{L^2} = \mathbb{E}(XY). \tag{160}$$

This space will be extremely important in the sequel and, for us, will be *the* example of a Hilbert space.

Given a Hilbert space, we define the length or *norm* of an element x by:

$$||x||_H := \sqrt{(x,x)_H}, \tag{161}$$

again often leaving off the subscript H. Clearly, $||\lambda \cdot x|| = |\lambda| \; ||x||$, if $\lambda \in \mathbb{R}$: multiplying x by the scalar λ means multiplying its length by the absolute value $|\lambda|$. If $H = \mathbb{R}^n$, then simply,

$$||x|| = \sqrt{x_1^2 + \cdots + x_n^2},$$

the Euclidean length of x. In the case of $L^2(\Omega, \mathcal{F}, \mathbb{P})$, the norm is:

$$||X||_{L^2} := \left(\int_\Omega X^2 d\mathbb{P} \right)^{1/2}. \tag{162}$$

This is often called the L^2*-norm.*

Another important concept is that of orthogonality: we say that $x, y \in$ are *orthogonal*, notation: $x \perp y$, if $(x,y) = 0$:

$$x \perp y \Leftrightarrow (x,y) = 0. \tag{163}$$

An important general fact about inner products is the so-called *Cauchy-Schwarz inequality*[19]:

$$|(x,y)| \leq ||x|| \cdot ||y||. \tag{164}$$

Another important general fact is the *triangle inequality:*

$$||x + y|| \leq ||x|| + ||y||.$$

To go from pre-Hilbert spaces to Hilbert spaces we have to explain the concept of completeness, which has to do with converging sequences in H. We say that a sequence $x_0, x_1, x_2, \cdots$ of elements of H *converges to an element* $x \in H$ if the norm of the difference goes to 0:

$$||x - x_n|| \to 0, \quad \text{as } n \to \infty \; .$$

We denote this by

$$x_n \to x \text{ in } H \; .$$

If $(x_n)_n$ is such a converging sequence, then the inter-distances $||x_n - x_m||$ will go to 0 as both m and n go to 0 simultaneously:

$$||x_n - x_m|| \to 0 \text{ as } n, m \to \infty. \tag{165}$$

Such sequences are called *Cauchy-sequences*, and what we just stated amounts to saying that converging sequences are Cauchy sequences.

[19]Russian mathematicians often add the name of *Buniatowski* who indeed seems to have been the first to discover it, presumably (as for the other two) in the context of $\mathbb{R}^n$

However, the *converse* of this statement need not be true, in general. Hilbert spaces are now precisely those pre-Hilbert spaces for which the converse *does* old:

Definition 8.3. A Hilbert space is a pre-Hilbert space for which every Cauchy-sequence has a limit.

This is called the *completeness property* of Hilbert spaces: it allows us to define elements of H by constructing Cauchy-sequences. A case in point will be the Ito integral. Two examples might help to clarify this:

Examples 8.4. (i) Not every Cauchy-sequence in $\mathbb{Q}$ converges in $\mathbb{Q}$: take for example a sequence of rational numbers converging to $\sqrt{2}$. This will be a Cauchy sequence, but its limit, $\sqrt{2}$, falls outside of $\mathbb{Q}$.

(ii)* Slightly more ambitiously, consider the space

$$V = \{f : [0,1] \to \mathbb{R} : f \text{ continuous}\},$$

with the L^2-inner product:

$$(f,g) = \int_0^1 f(x)g(x)dx.$$

This is a pre-Hilbert space, but not a Hilbert-space: consider a sequence of functions f_n, $n \geq 1$, which is:

1. 0 on $[0, \frac{1}{2} - \frac{1}{n}]$,
2. Linear on $[\frac{1}{2} - \frac{1}{n}, \frac{1}{2} + \frac{1}{n}]$, while 0 in the left end-point and 1 in te right end-point,
3. 1 on $[\frac{1}{2} + \frac{1}{n}, 1]$.

Then $(f_n)_n$ converges in the L^2-norm to the function which is 0 on $[0, \frac{1}{2})$, and 1 on $[\frac{1}{2}, 1]$ (it doesn't matter in fact what value we give it in the point $\frac{1}{2}$). It is therefore a Caucy sequence of elements in V, but the limit falls "outside of V".

A pre-Hilbert space can have "holes". There exists a general mathematical construction called *completion*, which amounts to "filling in all the holes" corresponding to non-converging Cauchy-sequences. Applied to the vector space V of the previous example, this would lead us to the space $L^2([0,1], \mathcal{B}([0,1]), dx)$. More generally, we have the following important theorem:

Theorem 8.5. *The space $L^2(\Omega, \mathcal{F}, \mathbb{P})$ is complete, and therefore a Hilbert space.*

We won't give the proof, which can be found in any text on measure theory. In practice, it is more important to be familiar with its statement, and to know and how and when to apply it.

8.2. **The projection theorem.** We now consider the following situation. Let H be a Hilbert space, and let $V \subset H$ be a subspace of H, that is, a subset V of H such that sums of elements are again in H, as are all products of elements of V by real numbers. Such a subspace is called *closed* is limits of converging sequences of elements of V are again in V, that is, if $x_n \in V$, $x_n \to x$ in H implies that $x \in V$ (observe that, a priori, we only know that x is in the larger space, H). An example of a subspace which is *not* close is the space V in example 8.4 (ii), regarded as a subspace of L^2.

Theorem 8.6. *(projection theorem) Let V be a closed subspace of the Hilbert space H. Then there exists, for each $h \in H$, a unique element $v \in V$ having smallest distance to h:*

$$||h - v|| = \min_{w \in V} ||h - w||. \tag{166}$$

The element $v \in V$ is called the projection of x onto V*, and is characterized by the property that $h - v$ is perpendicular to V, or*

$$(h - v, w) = 0 \text{ for all } w \in V. \tag{167}$$

We again skip the proof, as we will be more concerned with applying this theorem. For the finite dimensional Euclidean spaces $\mathbb{R}^n$ its statement should be relatively intuitive: make a drawing in $\mathbb{R}^3$ with V a plane, or a line). In infinite dimensions, some care is needed: example 8.4(ii) will again show that theorem 8.6 is false if V is not closed: see the exercises.

Given a closed subspace $V \subset H$, we define the *orthogonal projection* $P_V : H \to V$ by

$$P_V h = v \text{ if (166) (or equivalently, (167)) holds.} \tag{168}$$

Note that if we let $V^\perp := \{w \in H : (w, v) = 0 \text{ for all } v \in V \}$, then P_V maps $V^\perp$ onto 0. Further basic properties of $P = P_V$ are:

- (a) P_V is a projection, meaning that $P_V^2 = P_V$.
- (b) P_V is orthogonal, meaning that $(h - P_V h, v) = 0$, for all $v \in V$.

We note in passing that an equivalent way of stating (b) can be shown to be: for all $h, g \in H$: $(P_V h, g) = (h, P_V g)$.

8.3. **Application: Conditional expectations of finite variance rvs.** The following is an important application of this construction.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a σ-algebra, and let $\mathcal{G} \subset \Omega$ be some smaller σ-algebra, e.g. $\mathcal{G} = \sigma(X)$, the σ-algebra of information which can be gleaned from observing the random variable X. One can show that in this case the space

$$V = L^2(\Omega, \mathcal{G}, \mathbb{P})$$

of square-integrable $\mathcal{G}$-*measurable* rvs is a closed subspace of $L^2(\Omega, \mathcal{F}, \mathbb{P})$. Note that for a rv X to be in V, $X^{-1}((a,b))$ has to be in $\mathcal{G}$ instead of $\mathcal{F}$, which is a more stringent condition. The *conditional expectation with respect to* $\mathcal{G}$ can now be defined as being the orthogonal projection

$$\mathbb{E}(\cdot|\mathcal{G}) : L^2(\Omega, \mathcal{F}, \mathbb{P}) \to L^2(\Omega, \mathcal{G}, \mathbb{P}). \tag{169}$$

That is, if X is a rv with finite variance, then its *conditional expectation* $\Phi := \mathbb{E}(X|\mathcal{G})$ is characterized by the two following conditions:

- Φ is $\mathcal{G}$-measurable, and square-integrable: $\mathbb{E}(\Phi^2) < \infty$.
- For all square integrable and $\mathcal{G}$-measurable rvs $Y \in L^2(\Omega, \mathcal{G}, \mathbb{P})$:

$$\mathbb{E}(\Phi Y) = \mathbb{E}(XY), \tag{170}$$

since this is the same (in a different notation) as:

$$(X - \Phi, Y) = 0, \text{ for all } Y \in V = L^2(\Omega, \mathcal{G}, \mathbb{P}),$$

which characterized the orthogonal projection, according to the projection theorem.

One way to think about $\mathbb{E}(X|\mathcal{G})$ is that it represents the best prediction of the random variable X, given that we only dispose of the information $\mathcal{G}$. Indeed, remembering the equivalence of (166) and (167) in theorem 8.6, we may restate (170) as a variance-minimizing property (after substraction of the mean from X):

$$\text{Var } (X - \mathbb{E}(X|\mathcal{G})) = \min_{\substack{Y \ \mathcal{G}\text{-meas.} \\ \mathbb{E}(Y^2) < \infty}} \text{Var}\,(X - Y)\,. \tag{171}$$

We will return to conditional expectations in more detail (and from different points of view) later on.

8.4. **Exercises to chapter 8.**

Exercise 8.7. Let $H = L^2(\Omega, \mathcal{F}, \mathbb{P})$. Show, using the Cauchy-Schwarz-Buniatowski inequality (164), that if $X \in H$, then $\mathbb{E}(|X|) < \infty$. Hence the mean $\mathbb{E}(X)$ of X exists (is finite). Conclude that H is the same as the space of rvs on $(\Omega, \mathcal{F}, \mathbb{P})$ having finite mean and variance.

Exercise 8.8. Show by an example that theorem 8.6 is false if V is not closed.
(*Hint:* Example 8.4(ii), with H the space of L^2-integrable functions on $[0,1]$ and $h \in H$ the function which is 0 on $[0, 1/2)$, and 1 on $[1/2, 1]$.)

Exercise 8.9. Assuming the existence of a distance minimizing v, prove the equivalence of (166) and (167).

Exercise 8.10. Show that if $X \geq 0$ with probability 1, then the same holds for its conditional expectation $\mathbb{E}(X|\mathcal{G})$.
(*Hint:* a $\mathcal{G}$-measurable rv Φ is ≥ 0 with probability 1 iff $\mathbb{E}(\Phi \mathbb{I}_G) \geq 0$ for all $G \in \mathcal{G}$.)

9. The Ito stochastic integral

We will now return to the issue of defining a stochastic integral,

$$\int_0^T f_t dW_t, \tag{172}$$

where W_t is a Brownian motion. We now want to allow f_t to be stochastic also, and not just a deterministic function of time t, as in the discussion in chapter 3, and part of the problem is which stochastic f_t we can allow.

9.1. **Brownian motion revisited.**

With our new view of probability, Brownian motion will now consist of a family of random variables

$$W_t : \Omega \to \mathbb{R}$$

defined on some fixed probability space $(\Omega, \mathcal{F}, \mathbb{P})$ (for example, the one which we introduced in section 6.5, although other probability spaces are possible), and satisfying the usual axioms for Brownian motion:

- $W_0 = 0$,
- $W_t - W_s \sim N(0, t-s)$ for $s \leq t$,
- W_u, $W_t - W_s$ are independent if $u \leq s < t$.

The main advantage of this new view point is that we can now talk about the *sample paths* of Brownian motion, which are the functions:

$$t \to W_t(\omega) : \mathbb{R}_{\geq 0} \to \mathbb{R},\ \omega \in \Omega \text{ fixed.} \tag{173}$$

Note that there is a sample path for each $\omega \in \Omega$.

We can now also be more precise on the continuity of Brownian motion. In fact, we can (and will) suppose that $(\Omega, \mathcal{F}, \mathbb{P})$ and W_t are such that *with probability 1, the sample paths of W_t are continuous functions of t.* More precisely, there exists a null-set or null-event $N \subset \Omega$, $\mathbb{P}(N) = 0$, such that:

$$\text{for all } \omega \in \Omega \setminus N,\ t \to W_t(\omega) \text{ is continuous everywhere.} \tag{174}$$

It turns out that it is always possible to achieve this, by re-defining W_t suitably: cf. remark 6.18.

Theorem 9.1. *Brownian motion is almost surely nowhere differentiable. More precisely, there exists a null-event $N \in \mathcal{F}$ such that, for all $\omega \notin N$, the function $t \to W_t(\omega)$ is nowwhere differentiable (i.e., its graph does not have a tangent at any point).*

The proof consists again of some pretty technical mathematics; see for example Billingsly.

9.2. **Filtrations of σ-algebras.** Before starting the construction of the Ito-integral, we have to introduce an important new concept, that of a *(continuous) filtration of σ-algebras*. This is a family of σ-algebras $\mathcal{F}_t$, one for each $t \geq 0$, which is *increasing*, in the sense that:

$$s \leq t \Rightarrow \mathcal{F}_s \subset \mathcal{F}_t. \tag{175}$$

If we think of a σ-algebra as codifying information which can be obtained from making observations/doing statistical experiments, then we're simply dealing with an information set which grows with time. The following is a standard example:

Example 9.2. Let $(W_t)_{t\geq 0}$ be a Brownian motion, as above. Define $\mathcal{F}_0^W$ as being the collection of all null-sets in $\mathcal{F}$, and $\mathcal{F}_t^W$ as the σ-algebra generated by $\mathcal{F}_0^W$ together with all W_s for $s \leq t$:

$$\mathcal{F}_t^W = \sigma\left(\{W_s : s \leq t\} \cup \mathcal{F}_0^W\right). \tag{176}$$

Loosely speaking, $\mathcal{F}_t^W$ contains all possible information which can be obtained from observing Brownian motion up till, and including, time t (the null-sets counting as containing no information at all). The filtration $(\mathcal{F}_t^W)_{t\geq 0}$ is called the *Brownian filtration* (on the given probability-space $(\Omega, \mathcal{F}, \mathbb{P})$, to be precise).

Informally, $\mathcal{F}_t^W$-measurable functions can be thought of as functions of some, or all, of the W_u's with $u \leq t$: random variables Y of the form

$$Y = g(W_{u_1}, \cdots, W_{u_k}),$$

g a (for example) continuous function and $u_j \leq t$, are $\mathcal{F}_t^W$-measurable, and every $\mathcal{F}_t^W$-measurable Y can be shown to be a (point-wise) limit of a sequence of such special Y's. Now recall that, if $u \leq s < t$, then W_u and $W_t - W_s$ are independent. Functions of such W_u will also be independent of $W_t - W_s$, and, in view of the above, we obtain the following important property:

$$\begin{array}{c} \textit{If } u \leq s < t, \textit{ then every } \mathcal{F}_u^W\textit{-measurable} \\ \textit{function } Y \textit{ is independent of } W_t - W_s \end{array} \tag{177}$$

9.3. **Defining the Ito integral.** We now place ourselves in the following situation: besides our Brownian motion $(W_t)_{t\geq 0}$, we dispose of some filtration $(\mathcal{F}_t)_{t\geq 0}$ such that:

$$\text{Each } W_t \text{ is } \mathcal{F}_t\text{-measurable ,} \tag{178}$$

and such that the analogue of (177) holds:

$$\begin{array}{c} \text{If } u \leq s < t, \text{ then every } \mathcal{F}_u\text{-measurable} \\ \text{function } Y \text{ is independent of } W_t - W_s \end{array} \tag{179}$$

One can always take for $\mathcal{F}_t$ the Brownian filtration $\mathcal{F}_t^W$, but larger filtrations are also allowed, provided (179) is satisfied. This liberty of

choice is often convenient. For example, we might have *two* independent Brownian motions $(W_t)_t$ and $(Z_t)_t$. Then if

$$\mathcal{F}_t = \sigma\left(W_s, Z_r : s, r \leq t\right),$$

is the natural filtration generated by the two Brownian motions together, then $\mathcal{F}_t$ satisfies both (178) and (179), and is strictly bigger than $\mathcal{F}_t^W$, the filtration generated by W_t only.

The second ingredient for a stochastic integral is a stochastic process $H = (H_t)_{t\geq 0}$, $H_t : \Omega \to \mathbb{R}$, which will serve as the integrand in our stochastic integral (172), and which has to satisfy the following important condition:

$$H_t \text{ is } \mathcal{F}_t\text{-measurable, for each } t \geq 0\ . \tag{180}$$

We will say in this case that the process H_t is *adapted to the filtration* $\mathcal{F}_t$ *$(t \geq 0)$*, and also that it is *non-anticipating*[20].

For such $H = (H_t)_t$ we will now give a sense to the integral

$$I_T(H) = I_T(H)(\omega) = \int_0^T H_t(\omega) dW_t(\omega),\ \omega \in \Omega, \tag{181}$$

even though $t \to W_t(\omega)$ is, with probability 1, now where differentiable. Note that $I_T(H)$ will be a *function* on Ω, that is, a *random variable*.

We will construct the integral in two stages: first for a suitable class of *simple* adapted processes H_t, and then for more general H_t, by a process of approximation approximation and passing to the limit[21]. It is for this last step that the formalism of Hilbert spaces will prove to be very convenient.

9.4. **Ito's integral for simple adapted processes.** To define a simple process $(H_t)_{t\geq 0}$ we first need a finite partition

$$t_0 = 0 < t_1 < \cdots\ t_{N-1} = t_N = T \tag{182}$$

of the interval $[0, T]$: the points t_j divide the interval $(0, T]$ up in (typically small) subintervals $(t_{j-1}, t_j]$. As a concrete example you could think of

$$t_j = \frac{jT}{N},\ \ 0 \leq j \leq N,$$

where N is thought of as a big number. We next choose, for each t_j, an $\mathcal{F}_{t_{j-1}}$-measurable random variable K_j and call $(H_t)_{t\geq 0}$ *simple* if it is of the form:

$$H_t(\omega) = \sum_{j=1}^{n} K_j(\omega)\ \mathbb{I}_{(t_{j-1}, t_j]}(t), \tag{183}$$

[20]the origin of this term lies in the fact that when $\mathcal{F}_t = \mathcal{F}_t^W$, then such an H_t only depends on past to present values W_s, $s \leq t$, not on future values

[21]it is perhaps good to realize that *any* integral is always the result of some limit process, beginning with the basic integrals you were taught about in Calculus!

where we will put $H_0 = K_1$, by convention (its value in 0 won't matter much, in the present context). In other words:

$$H_t(\omega) = \begin{cases} K_1(\omega), \text{ if } 0 = t_0 < t \leq t_1, \\ K_2(\omega), \text{ if } t_1 < t \leq t_2, \\ \quad \vdots \\ K_n(\omega), \text{ if } t_{n-1} < t \leq t_n = T. \end{cases}$$

One can visualize such an H_t as a step-function on $[0, T]$, whose levels are given by the random numbers $K_j = K_j(\omega)$. Observe that H_t is adapted, since, for $t_{j-1} < t \leq t_j$, say,

$$H_t = K_j \text{ is } \mathcal{F}_{t_{j-1}}\text{-measurable, and therefore } \mathcal{F}_t\text{-measurable ,}$$

for $\mathcal{F}_{t_{j-1}} \subset \mathcal{F}_t$. We next define the Ito integral of such a simple function as:

$$\begin{aligned} I_T(H) &= \int_0^T H_t dW_t \\ &:= \sum_{j=1}^{N-1} H_{t_j}(\omega) \left(W_{t_{j+1}}(\omega) - W_{t_j}(\omega) \right) \end{aligned} \tag{184}$$

Observe that this is a function of ω, and therefore a random variable on Ω (although we usually suppress the variable ω when writing $\int_0^T H_t dW_t$); this is why this is called a *stochastic* integral.

The integral (184) has the following important properties:

Lemma 9.3. *If $(H_t)_{t \geq 0}$ is a simple adapted process, and if*

$$I_T = I_T(H) := \int_0^T H_t dW_t,$$

then:

(a) I_T has mean 0:

$$\mathbb{E}(I_T) = 0. \tag{185}$$

(b) I_T has variance

$$\mathbb{E}\left(I_T^2\right) = \int_0^T \mathbb{E}(H_t)^2 dt. \tag{186}$$

Proof of lemma 9.3: The proof of (a) is easy:

$$\begin{aligned} \mathbb{E}(I_T(H)) &= \sum_j \mathbb{E}\left(H_{t_j}(W_{t_{j+1}} - W_{t_j}) \right) \\ &= \sum_j \mathbb{E}(H_{t_j}) \mathbb{E}(W_{t_{j+1}} - W_{t_j}) \\ &\quad \text{(since } H_{t_j} \text{ and } W_{t_{j+1}} - W_{t_j} \text{ are independent)} \\ &= 0, \end{aligned}$$

since Brownian motion has mean 0.

The proof of (b) is a bit more involved:

$$\begin{aligned} I_T^2 &= \left(\sum_j H_{t_j}(W_{t_{j+1}} - W_{t_j})\right)^2 \\ &= \sum_j \sum_k \left(H_{t_j}(W_{t_{j+1}} - W_{t_j})\right)\left(H_{t_k}(W_{t_{k+1}} - W_{t_k})\right) \\ &= \sum_j H_{t_j}^2 (W_{t_{j+1}} - W_{t_j})^2 + \\ &\quad \sum\sum_{j\neq k} \left(H_{t_j}(W_{t_{j+1}} - W_{t_j}) H_{t_k}(W_{t_{k+1}} - W_{t_k})\right). \end{aligned}$$

We have to compute the expectation of this expression.

Now, by independence of the ($\mathcal{F}_{t_j}$-measurable) rv H_{t_j} and $(W_{t_{j+1}} - W_{t_j})$,

$$\begin{aligned} \mathbb{E}\left(H_{t_j}^2 (W_{t_{j+1}} - W_{t_j})^2\right) &= \mathbb{E}\left(H_{t_j}^2\right)\mathbb{E}\left((W_{t_{j+1}} - W_{t_j})^2\right) \\ &= \mathbb{E}(K_j^2)\cdot(t_{j+1} - t_j). \end{aligned}$$

As for the terms with $j \neq k$, if for example $j < k$, then $j+1 \leq k$, and

$$H_{t_j}(W_{t_{j+1}} - W_{t_j}) H_{t_k},$$

will be $\mathcal{F}_{t_k}$ measurable, and therefore independent of $W_{t_{k+1}} - W_{t_k}$. Hence,

$$\begin{aligned} & \mathbb{E}\left(H_{t_j}(W_{t_{j+1}} - W_{t_j}) H_{t_k}(W_{t_{k+1}} - W_{t_k})\right) \\ = \ & \mathbb{E}\left(H_{t_j}(W_{t_{j+1}} - W_{t_j}) H_{t_k}\right)\mathbb{E}\left(W_{t_{k+1}} - W_{t_k}\right) \\ = \ & 0. \end{aligned}$$

It follows that

$$\begin{aligned} \mathbb{E}(I_T^2) &= \sum_j \mathbb{E}(K_j^2)(t_{j+1} - t_j)^2 \\ &= \int_0^T \mathbb{E}(H_t^2)dt, \end{aligned}$$

since the function $t \to \mathbb{E}(H_t^2)$ is an ordinary step-function (with values in $\mathbb{R}$), which is equal to $\mathbb{E}(K_j^2)$ on $(t_{j-1}, t_j]$. QED

Remark 9.4. If we write out (186) in full, we get the statement:

$$\int_\Omega \left(\int_0^T H_t(\omega) dW_t(\omega)\right)^2 d\mathbb{P}(\omega) = \int_0^T \int_\Omega H_t^2(\omega) d\mathbb{P}(\omega) dt.$$

This is certainly not an obvious identity: when going from left to right we have to move the square under the second integral sign, to get it on $f_t(\omega)$, and this is false in general. To be able to do this, we needed

that H_t is adapted, and we also used the basic properties of Brownian motion, in particular independence of the future and the past-up-to-present. A closer analysis shows that the martingale property of Brownian motion, which amounts to the statement that

$$\mathbb{E}\left(W_t|\mathcal{F}_s\right)) = W_s,$$

is in fact all that is needed, and that one can generalize the stochastic integral by replacing $(W_t)_t$ by any square integrable martingale. We will precisely define and discuss martingales in a later chapter.

The identity (186) is fundamental for extending the Ito-integral to more general f_t's than just the simple ones.

9.5. **The Ito-integral in full generality.** The Ito-integral will now be extended from simple step-functions to more general ones by a limit argument. The basic idea is simply to approximate a general adapted integrand $H = (H_t)_t$ by a sequence of simple adapted integrands, $H_n = (H_{n,t})_t$:

$$H_{n,t}(\omega) \to H_t(\omega), \quad n \to \infty, \tag{187}$$

and to put:

$$I_T(H) = \lim_{n\to\infty} I_T(H_n) = \lim_{n\to\infty} \int_0^T H_{n,t}(\omega) dW_t(\omega), \tag{188}$$

hoping that this limit exists. Now the whole mathematical subtlety of the Ito integral lies in the way in which we have to interpret these two limits (187) and (188). Note that these are not simple limits of numbers, with which you should be familiar from elementary calculus, but limits of *functions*, namely functions on the sample space, Ω. The subject of limits in function spaces has been the object of intense mathematical research during the first half of the 20-th century, and has given rise to the field of Functional Analysis, and there exist several notions of convergence of sequences of functions. The Ito-integral is best understood in the context of Hilbert spaces; indeed, (186) can be understood as asserting that the Ito-integral on simple adapted functions is a length-preserving map, or isometry, between to suitably defined spaces of square integrable functions (L^2-spaces). We will therefore slightly change viewpoint and re-interpret everything in terms of functions on Ω and $[0,\infty) \times \Omega$, respectively. First observe that we can look upon a process $(H_t)_{t\leq T}$ as simply being a function

$$H : [0,T] \times \Omega \to \mathbb{R},$$

sending

$$H : (t,\omega) \to H_t(\omega).$$

We'll therefore also write $H_t(\omega)$ as $H(t,\omega)$. We now introduce the space of functions:

$$\mathcal{H}_T^2 = \{H : [0,T]\times\Omega \to \mathbb{R} \text{ adapted}, \int_0^T \int_\Omega H(t,\omega)^2 d\mathbb{P}(\omega)dt < \infty\}, \tag{189}$$

(where we'll tacitly assume that H is measurable with respect to the product σ-algebra $\mathcal{B}([0,T])\times\mathcal{F}$, in order to be able to give sense to the double integral; ignore this point if it confuses you).

If we introduce the inner product

$$(H,K)_{\mathcal{H}} = \int_0^T \int_\Omega H(t,\omega)K(t,\omega)dtd\mathbb{P}(\omega),$$

then $\mathcal{H}_T^2$ can be sown to be a Hilbert space[22]. The corresponding "infinite dimensional Euclidian length" is given by

$$||H||_{\mathcal{H}}^2 = \int_0^T \int_\Omega H(t,\omega)^2 d\mathbb{P}(\omega)dt,$$

which we can also write as:

$$||H||_{\mathcal{H}}^2 = \int_0^T \mathbb{E}(H_t)^2 dt.$$

$\mathcal{H}_T^2$ is the L^2-space of adapted processes. Recall form chapter 8 that there is also the (simpler) Hilbert space of random variables:

$$L^2(\Omega) = \{X : \Omega \to \mathbb{R} : \mathbb{E}(X^2) = \int_\Omega X(\omega)^2 d\mathbb{P}\omega) < \infty\},$$

with "Euclidian length"

$$||X||_{L^2}^2 = \int_\Omega X(\omega)^2 d\mathbb{P}(\omega) = \mathbb{E}(X^2).$$

We note in passing that using a more precise notation, $L^2(\Omega)$ should be designated as $L^2(\Omega,\mathcal{F},\mathbb{P})$, since it depends on both the σ-algebra and on the measure. A similar remark applies to $\mathcal{H}_T^2$, but we will keep to the simplified notations, so as not to make the following totally unreadable. It is clear from the previous that both $\mathcal{H}_T^2$ and $L^2(\Omega)$ are spaces of the same nature, namely functions whose square can be integrated, or are *square integrable*, with respect to a suitable measure.

One now can show the following important fact concerning $\mathcal{H}_T^2$:

Lemma 9.5. *Simple adapted processes (regarded as functions on* $[0,T]\times\Omega$ *are what is called* dense *in* $\mathcal{H}_T^2$*: for each* $H\in\mathcal{H}_T^2$ *there exists a sequence of simple functions* H_n*, such that*

$$||H - H_n||_{\mathcal{H}} \to 0, \text{ as } n\to\infty.$$

[22*] In fact, $\mathcal{H}_T^2$ is a closed subspace of the L^2-space $L^2([0,T]\times\Omega, \mathcal{B}([0,T])\times\mathcal{F}, dtd\mathbb{P})$

Equivalently, interpreting everything as processes, if $H = (H_t)_{0\leq t\leq T}$ is a stochastic process which is adapted to the filtration $\mathcal{F}_t, t \leq T$, then we can always find a sequence $H_n = (H_{n,t})_{0\leq t\leq T}$ of simple adapted processes, such that

$$||H - H_n||_{\mathcal{H}} = \int_0^T \mathbb{E}\left((H_t - H_{n,t})^2\right) dt \to 0. \tag{190}$$

The lemma is non-trivial, and requires a proof, for which we refer to the literature (cf. for example the book by Oksendael). Below we'll indicate how to construct such an approximating sequence if H has continuous paths.

We now show how to use these properties to define the Ito integral $I_T(H) = \int_0^T H_t dW_t$ of an arbitrary process $H = (H_t)_{t\geq 0}$ in $\mathcal{H}_T^2$.

Step 1. By the lemma, there exists a sequence of simple adapted processes, $H_n = (H_{n,t})_t$, such that

$$||H - H_n||_{\mathcal{H}} \to 0.$$

By the triangle inequality:

$$||H_n - H_m||_{\mathcal{H}} \leq ||H_n - H||_{\mathcal{H}} + ||H - H_m||_{\mathcal{H}} \to 0,$$

as $n, m \to \infty$ simultaneously.

Step 2. Being a simple adapted process, the Ito integral of H_n is already defined. We put

$$Y_n := I_T(H_n) = \int_0^T H_{n,t} dW_t.$$

Then $I_n - I_m = I_T(H_n - H_m)$, and (186) implies that:

$$\begin{aligned} ||Y_n - Y_m||^2_{L^2(\Omega)} &= \mathbb{E}\left((Y_n - Y_m)^2\right)) \\ &= \int_0^T \mathbb{E}\left((H_n - H_m)^2\right) dt \\ &= ||H_n - H_m||^2_{\mathcal{H}} \to 0, \end{aligned} \tag{191}$$

as $n, m \to \infty$ simultaneously. Hence:

$$(Y_n)_n \text{ is a Caucy-sequence in } L^2(\Omega) \text{ !}$$

Step 3. $L^2(\Omega)$, being a Hilbert space is complete, meaning that every Cauchy-sequence converges. There therefore exists an element $X \in L^2(\Omega)$ such that

$$||X - Y_n||^2_{L^2(\Omega)} \to 0,$$

that is,

$$\mathbb{E}\left((X - I_T(H_n))^2\right) \to 0,$$

and we define the stochastic integral of our original process H simply as being this limit:

$$I_T(H) = \int_0^T H_t dW_t := X = \lim_{n\to\infty} I_T(H_n).$$

Step 4. We finally have to show that this is a good definition, in the sense that the $I_T(H) := X$ we found does not depend on the approximating sequence $(H_n)_n$ of simple adapted processes (in general, there is more than one such sequence). But this is relatively easy: if $(\widehat{H}_n)_n$ is another such sequence, then

$$||H_n - \widehat{H}_n||_{\mathcal{H}} \to 0,$$

and therefore, using (186),

$$||I_T(H_n) - I_T(\widehat{H}_n)|| = ||H_n - \widehat{H}_n||_{\mathcal{H}} \to 0,$$

also. Hence, necessarily

$$\lim_{n\to\infty} I_T(H_n) = \lim_{n\to\infty} I_n(\widehat{H}_n).$$

This completes the construction of the Ito-integral.

It can be sown that $I_T()$ inherits the properties (185), (186) established the simple Ito integrals: the approximating $I_T(g_n)$ have these properties, and they persist in the limit (we won't give a formal proof). We record this formally as

Theorem 9.6. *Let $H = (H_t)_{0\leq t\leq T} \in \mathcal{H}_T^2$, and let*

$$I_T(H) = \int_0^T H_t dW_t,$$

its Ito-integral, which we just defined. Then

$$\mathbb{E}(I_T(H) = 0, \tag{192}$$

and

$$\mathbb{E}(I_T(H)^2) = \int_0^T \mathbb{E}(H_t^2)dt. \tag{193}$$

More generally, if $K = (K_t)_t$ is also in $\frac{2}{T}$, then

$$\mathbb{E}(\ I_T(H)\ I_T(K)) = \int_0^T \mathbb{E}(H_t K_t)dt. \tag{194}$$

The only point which perhaps needs comment is (194). This can either be established along the same lines as (193), by first verifying it for simple functions, and then passing to the limit (for which you'll need to use Cauchy's inequality, if you want to prove it rigorously). Alternatively, it can be derived from (193) by using a small trick called *polarization*: see the exercises at the end of this chapter.

9.6. **Ito's integral for integrands with continuous sample paths.** The above procedure is admittedly a bit abstract. However, if the integrand H_t has continuous sample paths $t \to H_t(\omega)$, and is bounded (which accounts for the vast majority of stochastic integrals used in Finance), we can give a less abstract and more concrete description of $I_T(H)$, which corresponds closely to the usual picture of integration in ordinary Calculus. The point is, that for such H_t we can choose a very simple type of approximating sequence of simple functions. For convenience, we will sometimes write

$$H(t,\omega) \text{ for } H_t(\omega).$$

We then put:

$$H_{n,t}(\omega) = \sum_{j=0}^{n-1} H\left(\frac{jT}{n},\omega\right) \mathbb{I}_{(jT/n,(j+1)T/n]}(t). \tag{195}$$

These are clearly simple adapted functions, since H_t is adapted. If the sample paths of H_t are continuous then, for each $\omega \in \Omega$ and $0 \le t \le T$:

$$H_{n,t}(\omega) \to H_t(\omega) \quad n \to \infty.$$

Under the additional condition that H *is bounded*, meaning that there exists a constant $C > 0$ such that

$$|H_t(\omega)| \le C, \quad \text{for all } \omega \in \Omega, 0 \le t \le T \ ,$$

one shows[23] that H_n tends to H in the space $\mathcal{H}_T^2$: $||H - H_n||_{\mathcal{H}} \to 0$

Next, if we compute the Ito-integral of $H_{n,t}$, then we obtain $I_T(H_{n,t}) = S_n(H,T)$, where:

$$S_n(H,T)(\omega) := \sum_{j=0}^{n-1} H\left(\frac{jT}{n},\omega\right) \big(W_{(j+1)T/n}(\omega) - W_{jT/n}(\omega)\big). \tag{196}$$

Observe that this is just like a Riemann sum, with the integrand H_t always evaluated in the *left* endpoint of the interval $[jT/N,\ (j+1)T/N]$. We then have:

Theorem 9.7. *If H_t is adapted and bounded, then*

$$\begin{aligned} I_T(H) &= \lim_{n\to\infty} S_n(H,T) \\ &= \lim_{n\to\infty} \sum_{j=0}^{n-1} H\left(\frac{jT}{n},\omega\right) \big(W_{(j+1)T/n}(\omega) - W_{jT/n}(\omega)\big) \ , \end{aligned} \tag{197}$$

in the sense that

$$\mathbb{E}\left(\left(I_T(H) - S_n(H,T)\right)^2\right) \to 0, \ \textit{as } n \to \infty \ . \tag{198}$$

[23] e.g. by applying Lebesgue's dominated convergence theorem

The convergence in (198) is called (for obvious reasons) *convergence in mean square sense.*

Observe that (197) immediately suggests how to compute numerically different realizations of $I_T(H)$, given sample paths of H_t and of W_t. Like for ordinary numerical integration, this simple algorithm can be improved: see the book by Kloeden and Platen.

9.7. **Ito processes.** Let $H = (H_t)_{t\geq 0}$ be an adapted processes, which is in $\mathcal{H}^2_t$ for each time $t \geq 0$:

$$\int_0^t \mathbb{E}(H_s^2)ds < \infty, \text{ all } t > 0$$

for each t. For such H the Ito-integrals

$$\int_0^t H_s(\omega)dW_s(\omega),$$

are well-defined and, as a function of the upper limit of integration t, defines a new process. Slightly more general, we can add an integral of the type

$$\int_0^t A_s(\omega)ds,$$

which are unproblematic if, for example, the process A_t has continuous paths: this is just an ordinary integral with respect to ds. Put

$$X_t = X_0 + \int_0^t A_s ds + \int_0^t H_s dW_s, \tag{199}$$

where X_0 is a constant; X_t is called an *Ito-process* and symbolically written as:

$$dX_t = A_t dt + H_t dW_t. \tag{200}$$

Remark on notation: To stress the fact that our integrands H_s in integrals like $\int_0^t H_s dW_s$ are allowed to be stochastic, we systematically designated them by capital letters in this chapter. However, in the end it becomes a bit tiresome to always use capital letters for random variables, and we will often revert also lower case letters h_t, a_t for stochastic processes, especially when they occur as integrands of Ito processes, in accordance with general notational practice in Stochastics and Finance. Ito processes are tus written as

$$dX_t = a_t dt + h_t dW_t,$$

etc., and it should be clear from the context whether a_t and h_t are stochastic or not.

In a rigorous mathematical development of the Ito-calculus, stochastic differentials like (200) are just a symbolic short-hand for the corresponding stochastic integral (199), and are not in themselves considered bona fide mathematical objects. However, the basic intuition obtained from manipulating stochastic differentials according to the Ito-rules we gave in chapter 3 is correct, and the results established using these rules can be rigorously proved after translating them into integrals. As an example we take another look at Ito's lemma.

9.8. * **Ito's lemma revisited.** For simplicity we will limit ourselves to the simplest and most basic form of Ito's lemma, formula (67): if $f = f(w)$ is 3-times continuously differentiable, with bounded derivatives, then:

$$df(W_t) = f'(W_t)dW_t + \frac{1}{2}f''(W_t)dt. \tag{201}$$

In integral form, this becomes:

$$\begin{aligned} f(W_t) - f(0) &= f(W_t) - f(W_0) \\ &= \int_0^t f'(W_s)dW_s + \int_0^t \frac{1}{2}f''(W_s)ds. \end{aligned} \tag{202}$$

**Proof.* We sketch a proof based on the construction of the Ito integral in the previous sections. The idea is to cut things up in small intervals again, and write:

$$f(W_t) - f(W_0) = \sum_{j=1}^{n-1} f\left(W_{(j+1)t/n}\right) - f\left(W_{jt/n}\right),$$

where both sides of the equation are of course functions of ω, which we suppress for legibility. One then uses the Taylor expansion of f to analyze each of the terms in the sum on the right, as follows:

$$\begin{aligned} f\left(W_{(j+1)t/n}\right) - f\left(W_{jt/n}\right) &= f'\left(W_{jt/n}\right)\left(W_{(j+1)t/n} - W_{jt/n}\right) \\ &+ \frac{1}{2}f''\left(W_{jt/n}\right)\left(W_{(j+1)t/n} - W_{jt/n}\right)^2 \\ &+ \text{Remainder } R_j \ . \end{aligned}$$

Since we assume that f is three times differentiable, with continuous and bounded derivatives, it follows from one of the standard Calculus formulas for Taylor with remainder, that the remainder term can be estimated by:

$$|R_j(w)| \leq C|W_{(j+1)t/n} - W_{jt/n}|^3, \tag{203}$$

where C is some sufficiently big constant which dominates the third derivative of f. Taking the sum over all j from 0 to $n-1$, proving

(202) then amounts to showing that the following limits hold, in mean square sense:

$$\sum_{j=0}^{n-1} f'\left(W_{jt/n}\right)\left(W_{(j+1)t/n}-W_{jt/n}\right) \to \int_0^t f'(W_s)dW_s, \tag{204}$$

$$\sum_{j=0}^{n-1} f''\left(W_{jt/n}\right)\left(W_{(j+1)t/n}-W_{jt/n}\right)^2 \to \int_0^t f''(W_s)ds, \tag{205}$$

and

$$\sum_{j=0}^{n-1} R_j \ \to \ 0. \tag{206}$$

To simplify the notations, we put

$$t_j = \frac{jt}{n}.$$

The first limit, (204), follows from theorem 9.7. To get some insight into the second, observe that the expectation of the left and side of (205),

$$\begin{aligned}\mathbb{E}\left(\sum_{j=0}^{n-1} f''\left(W_{t_j}\right)\left(W_{t_{j+1}}-W_{t_j}\right)^2\right) &= \sum_{j=0}^{n-1} f''\left(W_{t_j}\right)(t_{j+1}-t_j)\\ &=: J_n \to \int_0^t f''(W_s)ds,\end{aligned}$$

the last line by the definition of the ordinary (Riemann-) integral from Calculus. This is encouraging, but not yet enough, since we have to prove that the expectation of the square of the difference of the left hand side with the right hand side goes to 0. This is of course the same as saying that its square-root goes to 0, which is the L^2-norm. But since $||X||_{L^2} = \sqrt{\mathbb{E}(X^2)}$ satisfies the triangle inequality, it suffices to show that

$$||\sum_{j=0}^{n-1} f''\left(W_{t_j}\right)\left(W_{t_{j+1}}-W_{t_j}\right)^2 - J_n||_{L^2} \to 0, \tag{207}$$

for then

$$\begin{aligned} & || \sum_{j=0}^{n-1} f'' \left(W_{t_j}\right) \left(W_{(t_{j+1}} - W_{t_j}\right)^2 - \int_0^t f''(W_s) ds ||_{L^2} \\ \leq & || \sum_{j=0}^{n-1} f'' \left(W_{t_j}\right) \left(W_{t_{j+1}} - W_{t_j}\right)^2 - J_n ||_{L^2} \\ + & ||J_n - \int_0^t f''(W_s) ds ||_{L^2} \\ & \to 0, \end{aligned}$$

the first term on the right by (207), and the second by the definition of the ordinary integral[24].

So we are left with establishing (207). Expanding

$$\begin{gathered} \left(\left[\sum_{j=0}^{n-1} f''(W_{t_j}) \left(W_{t_{j+1}} - W_{t_j}\right)^2 \right] - J_n \right)^2 = \\ \left(\sum_j f''(W_{t_j}) (W_{t_{j+1}} - W_{t_j})^2 - (t_{j+1} - t_j) \right)^2 = \\ \sum_j \sum_k f''(W_{t_j}) f''(W_{t_k}) \left((W_{t_{j+1}} - W_{t_j})^2 - (t_{j+1} - t_j) \right) \\ \cdot \left((W_{t_{k+1}} - W_{t_k})^2 - (t_{k+1} - t_k) \right). \end{gathered}$$

We now take the expectation of all this, and examine the diagonal ($j = k$) and off-diagonal ($j \neq k$) separately. The last ones are easy since, if for example $j < k$, then by independence of future and past-to-present, their expectation equals

$$\mathbb{E}(\cdots)\mathbb{E}(\left(W_{t_{k+1}} - W_{t_k}\right)^2 - (t_{k+1} - t_k)) = 0,$$

since $W_{t_{k+1}} - W_{t_k}$ has variance $t_{k+1} - t_k$.

As regards the diagonal terms , we (again) expand:

$$\begin{gathered} \left((W_{t_{j+1}} - W_{t_j})^2 - (t_{j+1} - t_j) \right)^2 = \\ (W_{t_{j+1}} - W_{t_j})^4 - 2(t_{j+1} - t_j)(W_{t_{j+1}} - W_{t_j})^2 + (t_{j+1} - t_j)^2. \end{gathered}$$

Inserting this, and using again the independence of future increments, together with

$$\mathbb{E}(W_{t_{j+1}} - W_{t_j})^4) = 3(t_{j+1} - t_j)^3 = \frac{3t^2}{n^2},$$

and

$$\mathbb{E}((W_{t_{j+1}} - W_{t_j})^2) = t_{j+1} - t_j = \frac{t}{n},$$

[24]to be slightly more precise, by ordinary Calculus, for each ω, $J_n(\omega) \to \int_0^t f''(W_s(\omega)) ds$, for any $\omega \in \Omega$, and since everything is bounded, one can use Lebesgue's dominated convergence theorem to conclude that the L^2-norm of the difference also tends to 0

we finally find that

$$\begin{aligned} & | \mathbb{E}\left(\sum_{j=0}^{n-1} f''(W_{t_j})^2 \left((W_{t_{j+1}} - W_{t_j})^2 - (t_{j+1} - t_j)\right)^2\right) | \\ = & | \sum_{j=1}^{n-1} \frac{2t^2}{n^2} \mathbb{E}(f''(W_{t_j})^2 | \\ \leq & 2C' \cdot \frac{t^2}{n^2} \cdot n \simeq \frac{1}{n} \to 0, \end{aligned}$$

as $n \to \infty$, where we used that we can bound $\mathbb{E}(f''(W_{t_j})^2)$, uniformly for all j since, by hypothesis, the second derivative of f is bounded. This proves

Finally, similar arguments can be used to prove (206). This is in fact slightly easier, since by Cauchy-Schwarz and the bound (203),

$$\begin{aligned} \mathbb{E}\left(\left(\sum_j R_j\right)^2\right) & \leq n\mathbb{E}\left(\sum_j R_j^2\right)^2 \\ & \leq Cn \sum_{j=1}^{n-1} \mathbb{E}\left(W_{t_{j+1}} - W_{t_j})^6\right) \\ & = (Cn) \cdot (\frac{c_6 t^3}{n^3}) \cdot n \simeq \frac{1}{n} \to 0, \end{aligned}$$

where c_6 is the 6-t moment of the standard normal. This proves (206), and thereby the Ito formula in integral form (202) . QED

By using Taylors formula with integral remainder, one can weaken the hypothesis on f to twice continuously differentiable. One can also do away with the hypothesis that f'' be bounded; we refer to the literature for this.

More general forms of Ito's lemma, like for $dF(t, W_t)$, can be proved along the same lines (one now also has to expand to first order with respect to the t-variable).

9.9. **Exercises.**

Exercise 9.8. a) For any Hilbert space H, prove the *polarization identity*:

$$(x, y)_H = \frac{1}{2}\left(||x + y||_H^2 - ||x||_H^2 - ||y||_H^2\right),$$

$x, y \in H$.

b) Suppose we now have two Hilbert spaces H_1 and H_2, and a map $A : H_1 \to H_2$ such that 1) $A(x + y) = A(x) + A(y)$ and 2) $A(\lambda \cdot x) =$

$\lambda \cdot A(x)$. Here $x, y \in H$ and $\lambda \in \mathbb{R}$. Such maps are called *linear*. Suppose, moreover, that for all $x \in H_1$,

$$||A(x)||^2_{H_2} = ||x||^2_{H_1}$$

Show, using part a), that then

$$(A(x), A(y))_2 = (x, y)_{H_1},$$

for all $x, y \in H_1$. In words, *if a map between Hilbert spaces respects norms, then it respects inner products.*

c) Show that in theorem 9.6, (193) implies (194) .

10. Conditional Expectations and Martingales

10.1. **Conditional expectations.** To start, we take another look at conditional probabilities. You might want to first consult section 2.3, to refresh your memory: there we introduced conditional probability densities, for couples of random variables (X, Y) which posses a joint density. We will generalize this in two directions: first, we want to get away from the hypothesis that the random variables have densities, this sometimes being too restrictive in practice. Second, we will want to condition not only with respect to some other random variable, but with respect to a given *information set*, in the form of a sub-σ-algebra $\mathcal{G}$ of $\mathcal{F}$), if we're working in the context of a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

As a preliminary remark, it is good to realize that practically all quantities we're interested in, in Probability and Finance, can be expressed as an *expectation*. Even the probability of an event $F \in \mathcal{F}$ can be interpreted as the expectation of the indicator function $\mathbb{I}_F$ of that event, since:

$$\mathbb{P}(F) = \int_\Omega \mathbb{I}_F(\omega) d\mathbb{P}(\omega) = \mathbb{E}(\mathbb{I}_F). \tag{208}$$

We will study conditional expectations, in stead of just conditional probabilities.

Motivating Example: (You may skip this if you wish, an go directly to 10.1.) As a warm-up, we return to the conditional pdf's of section 2.3, and try to arrive at a formulation in terms of expectations instead of densities. So let X and Y be two rvs defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and having a joint density $f_{X,Y}(x, y)$, and put:

$$\mathbb{P}(X = x | Y = y) = \frac{f_{X,Y}(x, y)}{f_Y(y)},$$

assuming more-over that the denominator is non-zero. Let $g : \mathbb{R} \to \mathbb{R}$ be a "reasonable" function[25], and let us compute the expectation of $Xg(Y)$:

$$\begin{aligned}
\mathbb{E}(Xg(Y)) &= \int\int_{\mathbb{R}\times\mathbb{R}} xg(y) F_{X,Y}(x, y) dxdy \\
&= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \frac{f_{X,Y}(x, y)}{f_Y(y)} dx \right) g(y) f_Y(y) dy \\
&= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} x \mathbb{P}(X = x | Y = y) dx \right) f_Y(y) dy.
\end{aligned}$$

Now we will give this expression a new twist: introduce a new function $\varphi : \mathbb{R} \to \mathbb{R}$ by:

$$\varphi(y) = \int_{\mathbb{R}} x \mathbb{E}(X = x | Y = y) dx,$$

[25] Borel measurable, say

and a new random variable $\Phi : \Omega \to \Omega$ by:

$$\Phi(\omega) = \varphi(Y(\omega)).$$

Then, clearly, by the previous computation,

$$\begin{aligned} \mathbb{E}(\Phi g(Y)) &= \mathbb{E}(\varphi(Y)g(Y)) \\ &= \int_{\mathbb{R}} \varphi(y)g(y)f_Y(y)dy \\ &= \mathbb{E}(Xg(Y)). \end{aligned}$$

You may well wonder what the purpose is of this computation, and ask what we have gained by writing $\mathbb{E}(Xg(Y))$ as $\mathbb{E}(\Phi g(Y))$. What we have gained is, that Φ, being equal to a function $\varphi(Y)$ of Y, *contains the same information as* Y, while X, in general, does not. Let us formalize this later point by introducing

$$\sigma(Y),$$

the *σ-algebra generated by* Y, cf. formula (144) in section 6.5. It is a theorem that $\varphi(Y)$ is $\sigma(Y)$-measurable, since functions of Y are σ_Y-measurable, if the function in question is reasonably "nice" (e.g. Borel measurable itself), and our function φ above is sufficiently nice. Hence:

$$\Phi \text{ is } \sigma(Y)\text{-measurable.} \tag{209}$$

Furthermore, since φ is independent of the function $g = g(y)$, we clearly have that:

$$\mathbb{E}(\Phi g(Y)) = \mathbb{E}(Xg(Y)).$$

One traditionally formulates this last point a little differently. One can proof[26] that any $\sigma(Y)$-measurable function can be written as some function $g(Y)$ of Y. Applying this to the indicator function $\mathbb{I}_G$ of an arbitrary $G \in \sigma(Y)$, we find that:

$$\mathbb{E}(\Phi \mathbb{I}_G) = \mathbb{E}(X\mathbb{I}_G), \quad \text{for any } G \in \sigma(Y)\ . \tag{210}$$

This last equation can also be written as:

$$\int_G \Phi(\omega)d\mathbb{P}(\omega) = \int X(\omega)d\mathbb{P}(\omega). \tag{211}$$

The new random variable Φ is clearly determined a.e. by (209) and (210). It is called the conditional expectation of X with respect to $\sigma(Y)$, and is denoted by

$$\Phi = \mathbb{E}(X|\sigma(Y)),$$

and often also, more sloppily, by

$$\mathbb{E}(X|Y).$$

We observe the following two important points:

[26] this is actually a theorem of J. Doob

- $\mathbb{E}(X|\sigma(Y))$ is a *random variable*, that is, a *function* (on Ω), and *not* a number: this may take some getting used to.
- The point about $\Phi = \mathbb{E}(X|\sigma(Y))$ is not only that it satisfies (210), (211), but *that it does that in conjunction with being* $\sigma(Y)$*-measurable*, (209). Indeed, (210) by itself is already satisfied by X itself, and this condition on its own is rather empty.

We now give the general definition of conditional expectation with respect to a sub σ algebra of $\mathcal{F}$:

Definition and Theorem 10.1. *Given a random variable X on $(\Omega, \mathcal{F}, \mathbb{P})$ and a (smaller) σ-algebra $\mathcal{G} \subset \mathcal{F}$, there exists another random variable $\Phi : \Omega \to \mathbb{R}$ satisfying:*

(i) Φ is $\mathcal{G}$-measurable: events of the form

$$\{\omega : a < \Phi(\omega) < b\} \quad (a, b \in \mathbb{R})$$

are already in $\mathcal{G}$ instead of just in $\mathcal{F}$.

(ii) For all $G \in \mathcal{G}$: $\mathbb{E}(X\mathbb{I}_G) = \mathbb{E}(\Phi\mathbb{I}_G)$.

This new rv Φ is essentially uniquely determined in the sense that another rv Φ' satisfying (i) and (ii) will only differ from Φ on a null-set. We will simple ignore such differences, and treat Φ as being unique. We will write:

$$\Phi = \mathbb{E}(X|\mathcal{G}), \tag{212}$$

and call this the conditional expectation of X with respect to $\mathcal{G}$.

The proof of this theorem uses a deep result from abstract measure theory called the *Radon - Nikodym theorem*, which is outside the scope of these lectures, and we will just assume the existence of such a $\Phi = \mathbb{E}(X|\mathcal{G})$[27]. We stress, as before, that $\mathbb{E}(X|\mathcal{G})$ is a *random variable* and not a number, and that it is the *combination* of (i) and (ii) which makes it interesting, and which causes it to be (essentially) unique.

Rephrasing the above definition, we see that $\mathbb{E}(X|\mathcal{G})$ is uniquely determined by the following two properties:

$$\mathbb{E}(X|\mathcal{G}) \quad \mathcal{F} - \text{measurable}, \tag{213}$$

and

$$\mathbb{E}(X\mathbb{I}_G) = \mathbb{E}\left(\mathbb{E}(X|\mathcal{G}) \cdot \mathbb{I}_G\right), \quad \text{all } G \in \mathcal{G}\ . \tag{214}$$

Another way of stating these last equations is:

$$\int_G X(\omega) d\mathbb{P}(\omega) = \int_G \mathbb{E}(X|\mathcal{G})(\omega) d\mathbb{P}(\omega), \tag{215}$$

[27] *there is an alternative way to define $\mathbb{E}(X|\mathcal{G})$ for rvs X for which $\mathbb{E}(X^2) < \infty$, using Hilbert space theory: one uses the inner product $(X, Y) = \mathbb{E}(XY)$, and defines $\mathbb{E}(X|\mathcal{G})$ as the orthogonal projection onto the subspace of $\mathcal{G}$–measurable rvs; see also proposition 10.3 below

for all $G \in \mathcal{G}$. Whichever of these two formulations you prefer is largely a matter of taste.

It follows from (214) with $G = \Omega$ (which is an element of $\mathcal{G}$ since it is an element of any σ-algebra) that X and $\mathbb{E}(X|\mathcal{G})$ have the same mean:

$$\mathbb{E}\left(\mathbb{E}(X|\mathcal{G})\right) = \mathbb{E}(X). \tag{216}$$

We next list some other properties of conditional expectations which are important for computations, especially (iii) and (iv) below.

Proposition 10.2. *(i) Taking conditional expectations is linear:*

$$\mathbb{E}(X_1 + X_2|\mathcal{G}) = \mathbb{E}(X_1|\mathcal{G}) + \mathbb{E}(X_2|\mathcal{G})$$

and, if $\lambda \in \mathbb{R}$,

$$\mathbb{E}(\lambda X|\mathcal{G}) = \lambda \mathbb{E}(X_1|\mathcal{G}).$$

(ii) The trivial conditional expectation: *if* $\mathcal{G} = \mathcal{F}_{triv} = \{\emptyset, \Omega\}$, *the trivial sigma-algebra, then we simply have that*

$$\mathbb{E}(X|\mathcal{F}_{triv}) = \mathbb{E}(X),$$

the ordinary expectation.

(iii) If Y *is* $\mathcal{G}$*-measurable, then*

$$\mathbb{E}(Y|\mathcal{G}) = Y;$$

more generally, for any X,

$$\mathbb{E}(XY|\mathcal{G}) = Y\mathbb{E}(X|\mathcal{G}).$$

(iv) The tower property: *if* $\mathcal{H} \subset \mathcal{G} \subset \mathcal{F}$ *is a ascending chain of* σ*-algebras, and if* X *is a (*$\mathcal{F}$*-measurable) random variable, then:*

$$\mathbb{E}(X|\mathcal{H}) = \mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{H}).$$

**Proof.* The proof basically consists in verifying that the right hand side of the equations satisfies the defining properties of the conditional expectation on the left. For example, for the tower property, this goes as follows:

- $\mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{H})$, being the conditional expectation of the rv $\mathbb{E}(X|\mathcal{G})$ with respect to $\mathcal{H}$ is, by definition, $\mathcal{H}$-measurable, and therefore satisfies defining property (213).
- Next, if $H \in \mathcal{H}$, then since $\mathcal{H} \subset \mathcal{G}$, we also have that $H \in \mathcal{G}$, and therefore, by definition,

$$\mathbb{E}(X\mathbb{I}_H) = \mathbb{E}\left(\mathbb{E}(X|\mathcal{G})\mathbb{I}_H\right). \tag{217}$$

  Putting momentarily $Z = \mathbb{E}(X|\mathcal{G})$, we have, once more by definition, but now of conditional expectation with respect to $\mathcal{H}$:

$$\mathbb{E}(Z\mathbb{I}_H) = \mathbb{E}\left(\mathbb{E}(Z|\mathcal{H}) \cdot \mathbb{I}_H\right).$$

Substituting what Z is, we see that $\mathbb{E}(\mathbb{E}|\mathcal{G}|\mathcal{H})$ satisfies property (214) of $\mathbb{E}(X|\mathcal{H})$.

The verification of the other properties is similar. $$

The next proposition gives a different interpretation of the conditional expectation: it shows that if we try to "predict" the rv X (which is $\mathcal{F}$-measurable) by random variables Y which "only contain information relative to the smaller σ-algebra $\mathcal{G}$" (are only $\mathcal{G}$-measurable), then $\mathbb{E}(X|\mathcal{G})$ is that one among the Y's such that the variance of $X-Y$ is smallest:

Proposition 10.3. *(The* variance minimizing *property)*

$$\mathrm{Var}\,(X-\mathbb{E}(X|\mathcal{G})) = \min_{Z\ \mathcal{G}-measurable} \mathrm{Var}(X-Z),$$

where X and Z are supposed to have finite variance. This is in fact equivalent to the following property, which one may call the orthogonality property:

$$\mathbb{E}\,(X-\mathbb{E}(X|\mathcal{G})|\mathcal{G}) = 0.$$

We will skip the proof, although it is not that difficult: its main idea is the same as the proof of simpler result you may be familiar with, namely that the least-squares method in statistics gives the variance-minimizing prediction of a line through a set of points in the plane.

We next turn to the relation between conditional expectation and independence.Recall. from chapter 2, that if X and Y are independent, then: $\mathbb{P}(X=x|Y=y)=\mathbb{P}(X=x)$ and, consequently,

$$\begin{aligned}\mathbb{E}(X|Y=y) &= \int_{\mathbb{R}} x\mathbb{P}(X=x|Y=y)dx\\ &= \int_{\mathbb{R}} x\mathbb{P}(X=x)dx\\ &= \mathbb{E}(X),\end{aligned}$$

assuming everything in sight has a density, of course. We now want to generalize this to conditional expectations with respect to σ-algebras.

First, we say that a rv X is *independent of the σ-algebra $\mathcal{G}$* if, for all functions $g:\mathbb{R}\to\mathbb{R}$ (which are Borel measurable, say):

$$\mathbb{E}(g(X)\mathbb{I}_G) = \mathbb{E}(g(X))\mathbb{E}(\mathbb{I}_G)(=\mathbb{E}(g(X))\mathbb{P}(G)). \tag{218}$$

Equivalent ways of stating this are:

$$\mathbb{E}(g(X)Z) = \mathbb{E}(g(X))\mathbb{E}(Z),\ \text{for all } \mathcal{G}\text{-measurable } Z\ , \tag{219}$$

or

$$\mathbb{P}(A\cap G) = \mathbb{P}(A)\mathbb{P}(G),\ \text{for all } G\in\mathcal{G}, A\in\mathcal{F}_X\ . \tag{220}$$

Note that the latter equation is the same as:

$$\mathbb{E}(\mathbb{I}_A\mathbb{I}_G) = \mathbb{E}(\mathbb{I}_A)\mathbb{E}(\mathbb{I}_G).$$

With this notion in place, we can now state the following result:

Proposition 10.4. *If X is independent of the σ-algebra $\mathcal{G}$, then its conditional expectation is simply the ordinary expectation (and thus a number):*

$$\mathbb{E}(X|\mathcal{G}) = \mathbb{E}(X).$$

Proof. We just have to check that $\mathbb{E}(X)$ (or, more precisely, the function on Ω which is constantly equal to $\mathbb{E}(X)$) satisfies the defining properties (i) and (ii) of definition 10.1. But this is easy: constant functions are $\mathcal{F}_{triv}$-measurable, so certainly $\mathcal{G}$-measurable, since any σ-algebra always contains the trivial one. Next, if $G \in \mathcal{G}$, then since X is independent of $\mathbb{I}_G$,

$$\mathbb{E}(X\mathbb{I}_G) = \mathbb{E}(X)\mathbb{P}(G) == \mathbb{E}\left(\mathbb{E}(X) \cdot \mathbb{I}_G\right),$$

for $\mathbb{E}(X)$ is just a constant which can be pulled in front of the expectation sign, in the last equality.

This shows that $\mathbb{E}(X)$ satisfies conditions (i) and (ii) of 10.1, which completes the proof. $$

10.2. **Martingales.** Intuitively, one should look at conditional expectation $\mathbb{E}(X|\mathcal{G})$ as the "optimal prediction of X, given the information $\mathcal{G}$". This leads quite naturally to the idea of a martingale. Suppose we have an increasing family of σ-algebras $\mathcal{F}_t \subset \mathcal{F}, t \geq 0$:

$$\mathcal{F}_s \subset \mathcal{F}_t, \quad s < t,$$

and that for each $t \geq 0$ we have a random variable X_t which is $\mathcal{F}_t$-measurable; in the language of the previous chapter, $(X_t)_{t\geq 0}$ is a *process which is adapted to the filtration* $(\mathcal{F}_t)_{t\geq 0}$. We say that $(X_t)_{t\geq 0}$ *is a martingale with respect to* $(\mathcal{F}_t)_{t\geq 0}$ if, for all $s < t$,

$$\mathbb{E}(X_t|\mathcal{F}_s) = X_s, \tag{221}$$

almost everywhere as functions on Ω (with respect to the given probability). The idea is that X_t is a random-variable such that the best prediction of any of its future values, given today's information, is today's value; one may think of the weather, or of IBM's stock-price.

More precisely, $(X_t)_{t\geq 0}$ is called a *continuous-time martingale*; a *discrete-time martingale* is simply one in which the index t runs over a discrete set (like $\mathbb{N} = \{0, 1, 2, \cdots\}$).

Remark 10.5. It is **quite important** to realize that *in the definition of a martingale, the filtration $\mathcal{F}_t$ plays as big a rôle as the rvs X_t*: it is quite easily possible that X_t ceases to be a martingale of we change the filtration (like, replacing it by a bigger one), simply because *conditional expectation depends on the σ-algebra with respect to which we condition.* In much of the economics literature one simply writes

$$\mathbb{E}_t(X) \text{ for } \mathbb{E}(X|\mathcal{F}_t), \tag{222}$$

specifying $\mathcal{F}_t$ vaguely as "all information available at time t). Although we will sometimes also indulge in (222) (mainly since it saves time and typing-effort) it is actually very bad notation, if one does not clearly specify the context (that is, the $\mathcal{F}_t$ one uses). One can easily imagine situations in Finance in which the information set makes a difference, e.g. when there is insider training.

A basic example of a martingale is Brownian motion: we will check this carefully in the next example.

Example 10.6. Let $(W_t)_{t\geq 0}$ be a Brownian motion on our given probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If $s < t$ then, since $W_t - W_s$ is independent of $\mathcal{F}_s^W$ (which, we recall, is the σ-algebra generated by all W_u's for $u \leq s$), we have that:

$$\begin{aligned}
\mathbb{E}(B_t|\mathcal{F}_s^W) &= \mathbb{E}(W_s + (W_t - W_s)|\mathcal{F}_s^W) \\
&= \mathbb{E}(W_s|\mathcal{F}_s^W) + \mathbb{E}(W_t - W_s|\mathcal{F}_s^W) \ \text{ (by prop. 10.2))} \\
&= W_s + \mathbb{E}(W_t - W_s) \ \text{ (by prop. 10.2 (ii) and prop. ??, resp.)} \\
&= W_s \ \text{ (since } W_t - W_s \text{ has mean 0)}.
\end{aligned}$$

Hence $(W_t)_{t\geq 0}$ is a martingale with respect to the filtration $\mathcal{F}_t^W$.

The processes introduced in this exercise are called *Lévy-processes*. Brownian motion is of course an example of such a process, but there are many others, including manu which are much more irregular than Brownian motion, having for example everywhere discontinuous sample paths.

10.3. **Martingales and Ito-integrals.** There are two important results here.

Theorem 10.7. *Let $(\mathcal{F}_t)_{t\geq 0}$ be a filtration satisfying conditions (178) and (179) in subsection 8.3, and let $f = (f_t)_{t\geq 0}$ be an adapted process as in theorem 9.7 ([28], so that the Ito-integrals*

$$I_t = \int_0^t f_s dW_s,$$

are all well-defined. Then $(I_t)_{t\geq 0})$ is a martingale with respect to the given filtration $\mathcal{F}_t)_{t\geq 0}$.

The idea of the proof is rather straightforward: one first checks the theorem in case the process of the f_t's is simple, for which the result is practically true by definition, and one then derives it in general by taking limits. As usual with these matters, we skip the details, and refer to the literature.

In some sense, every martingale is an Ito-integral, provided it is a martingale with respect to a Brownian filtration. The precise formulation is as follows:

[28] or, more generally, let $f \in \mathcal{H}_t^2$ for all $t \geq 0$; cf. (189)

Theorem 10.8. *Let $(X_t)_{t\geq 0}$ be a martingale with respect to a Brownian filtration $(\mathcal{F}_t^W)_{t\geq 0}$ (for a given Brownian motion $(W_t)_{t\geq 0}$ on our probability space), such that the second moments $\mathbb{E}(X_t^2)$ stay bounded on each bounded time-interval $[0,T]$. Then X_t can be written as an Ito-integral: there exists an adapted process $(f_t)_{t\geq 0}$ such that with probability 1 on Ω,*

$$X_t(\omega) = X_0 + \int_0^t f_s(\omega) dW_s(\omega),$$

The proof of this theorem is considerably more complicated; it is also *non-constructive*, in the sense that it does not provide an explicit description of f_t, but only shows that there exists one (based, in the final analysis,on an argument by contradiction). Nevertheless, it plays an important rôle in the martingale approach to Asset Pricing, as will be explained in Pricing II.

10.4. **Exercises to Chapters 9 and 10.**

Exercise 10.9. Which of the following stochastic integrals make sense?

(a)
$$\int_0^1 (W_t^2 + 2W_{t/2} + 1) dW_t.$$

(b)
$$\int_0^1 W_{2t} dW_t.$$

(c)
$$\int_1^\infty W_{1/t} dW_t.$$

(d)
$$\int_0^1 W_{1/t} dW_t$$

(e)
$$\int_0^1 W_{2t} dt + \int_0^1 W_{t/2} dW_t.$$

(f)
$$\int_0^1 \frac{1}{\sqrt{W_t}} dW_t.$$

(g)
$$\int_0^1 \frac{1}{W_t^a} dW_t,$$

for $a < 1/2$.

(h)
$$\int_0^1 \frac{1}{t^{1/4} W_t^{1/4}} dW_t.$$

Exercise 10.10. Argue, that the following repeated stochastic integral is well-defined:

$$\int_0^t \left(\int_0^u dW_u \right) dW_v.$$

Evaluate this integral.

Exercise 10.11. Show that

$$\int_0^t W_s^2 dW_s = \frac{1}{3} W_t^3 - \int_0^1 W_s ds.$$

For the next two exercises you might need the following result from Probability Theory:

Theorem: *If Y_n is a sequence of Gaussian rv, and if $Y_n \to Y$ in the sense that $\mathbb{E}\,(Y_n - Y)^2) \to 0$, then the limit Y is also Gaussian.*

This can in fact be proved using characteristic functions.

Exercise 10.12. This question studies the integral

$$X_t = \int_0^t W_s ds.$$

(a) Argue that X_t is a Gaussian random variable.

(b) Compute it's mean and variance.

Exercise 10.13. Let $g = g(t)$ be a function of t only, which is for example continuous (or, as a weaker condition, for which $\int_0^t g(s)^2 ds$ stays finite, for all t). Define I_t by:

$$I_t = \int_0^t g(s) dW_s.$$

(a) Show that I_t is a Gaussian variable of mean 0, and variance:

$$f(t) = \int_0^t g(s)^2 ds.$$

Now suppose that $g(s) \neq 0$ a.e., so that the function $t \to \sigma(t)^2$ is strictly increasing, and therefore invertible.

We now define a new process, $\widetilde{W}_u$, by:

$$\widetilde{W}_u = X_t \text{ if } u = f(t).$$

(b) Prove that $\widetilde{W}_u)_{u \geq 0}$ is a (new) Brownian motion.

Significance of this exercise: Clearly,

$$X_t = \widetilde{W}_{f(t)},$$

and we say that X_t is obtained from the Brownian motion $\widetilde{W}_t$ by a *time-change* $t \to f(t)$. This can be generalized to stochastic, adapted, integrands $g_t = g(t, \omega)$, and to solutions of SDE's.

Exercise 10.14. Take $\Omega = \mathbb{R}$, $\mathcal{F} = \mathcal{B}(\mathbb{R})$, and $\mathbb{P}$ a probability-measure defined by a probability-density f, via:

$$\mathbb{P}(F) = \int_F f(x)dx, \quad F \in \mathbb{R} \text{ qa Borel set.}$$

In particular,

$$\mathbb{P}\left((a,b]\right) = \int_a^b f(x)dx.$$

(a) Let $\mathcal{G}$ be the σ-algebra generated by all subintervals $(n, n+1]$, $n \in \mathbb{N}$ (draw these on the line!). Check that $\mathcal{G}$ is exactly the collection of all unions of such intervals.

(b) Proof that $X : \Omega = \mathbb{R} \to \mathbb{R}$ is $\mathcal{G}$-measurable precisely when X is constant on each of the intervals $(n, n+1]$ (basically because these are the smallest elements of $\mathcal{G}$).

(c) Now let $X : \mathbb{R} \to \mathbb{R}$ be an arbitrary random variable, that is, an arbitrary $\mathcal{B}(\mathbb{R})$-measurable function. Show that $\mathbb{X}|\mathcal{G})$ is given by:

$$\mathbb{E}(X|\mathcal{G})(x) = \frac{1}{\mathbb{P}\left((n,n+1]\right)} \int_n^{n+1} X(y)f(y)dy, \text{ if } x \in (n, n+1] \ ;$$

that is, $\mathbb{E}(X|\mathcal{G})$ is, on $(n, n+1]$, constantly equal to the average (with respect to the given probability-measure) of X on $(n, n+1]$.

Exercise 10.15. Suppose that $(X_t)_{t\geq 0}$ is a stochastic process such that:

- X_t has mean 0: $\mathbb{E}(X_t) = 0$,
- X_t has what is called *independent increments*, meaning that if $u \leq s < t$, then $X_t - X_s$ is independent of X_u.

Let $\mathcal{F}_t^X$ be the filtration generated by the process X_t:

$$\mathcal{F}_t^X = \mathcal{F}\left(X_u^{-1}((a,b)) : u \leq t, a, b \in \mathbb{R}\right).$$

(Compare the definition of a Brownian filtration.) Show that $X_t, t \geq 0$ is a martingale with respect to $\mathcal{F}_t^X$.
(*Hint:* copy the argument of the example 10.6.)

Exercise 10.16. Consider an Ito-process:

$$dX_t = a_t dt + \sigma_t dW_t.$$

(a) Show that X_t will not be a martingale (w.r.t. the natural Brownian filtration), unless $a_t = 0$, for all t.

(b) Define the new process M_t by:

$$M_t = \exp\left(-\int_0^t h_s dW_s - \frac{1}{2}\int_0^t h_t^2 dt\right).$$

Show that M_t is a martingale.

(*Hint:* write M_t as $M_t = \exp(Y_t)$ for a suitable process Y_t, and compute dM_t using Ito's lemma; show that $dM_t = -h_t M_t dW_t$, and use a result from the course).

(c) Show that if we choose $h_t = a_t/\sigma_t$, then $X_t M_t$ is also a martingale, by proving that:

$$d(X_t M_t) = (\sigma_t - h_t X_t) M_t dW_t.$$

So we can change Ito-processes which aren't martingales into ones which are! This is intimately connected with Girsanov's theorem and risk-neutral pricing, as we will see in Pricing II. The next exercise gives the flavor:

Exercise 10.17. Let

$$dS_t = \mu S_t dt + \sigma S_t dW_t.$$

be a geometric Brownian motion representing for example an asset price.

(a) Let $Z_t = e^{-rt} S_t$, the discounted price process. Determine an SDE for Z_t.

(b) Find a martingale process M_t such that $Z_t M_t$ is also a martingale.

11. Feynman-Kac revisited

We will now prove the general case of the Feynman-Kac formula, for a partial differential equation whose 0-th order term may be non-constant. The general type of partial differential equation one encounters in derivative pricing has the following form:

$$\frac{\partial V}{\partial t} + \frac{\sigma(x,t)^2}{2}\frac{\partial^2 V}{\partial x^2} + a(x,t)\frac{\partial V}{\partial x} - c(x,t)V = 0. \tag{223}$$

We wish to solve this equation for $t < T$, with a (final) boundary condition :

$$V(x,T) = F(x), \tag{224}$$

where $F(x)$ is a given function.

Example 11.1. The Black and Scholes valuation equation, where, as usual, we let "S" designate the independent variable, instead of "x" (S represents the risky asset's value on which the option is written) :

$$\frac{\partial V}{\partial t} + \frac{\sigma^2 S^2}{2}\frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial x} - rV = 0. \tag{225}$$

We see that:

- $a(S,t) = rS$, r a constant,
- $\sigma(S,t) = \sigma S$, σ a constant,
- $c(S,t) = r$, constant,

and that $V(S,T) = F(S)$ is the final pay-off:

- $F(S) = \max(S - X, 0)$ for a European call with strike X.
- $F(S) = \max(X - S, 0)$ for a European put with strike X.

Example 11.2. The bond-pricing equation for a maturity T-bond in the Vasicek model for the short interest rate is of the form:

$$\frac{\partial P}{\partial t} + \frac{\sigma^2}{2}\frac{\partial^2 P}{\partial r^2} + \alpha(\theta - r)\frac{\partial P}{\partial r} - rP = 0. \tag{226}$$

Here $P = P(r,t)$ is a function of time t and of the short rate r (which now is the *variable*, and not a constant as in the previous example), and we are looking for a solution for time $t < T$ satisfying, at $t = T$, the boundary condition

$$P(r,T) = 1,$$

the promised pay-off of the bond at maturity. We see that now (replacing "x" by "r "):

- $a(r,t) = \alpha(\theta - r)$,
- $\sigma(r,t) = \sigma$, a constant,
- $c(r,t) = r$, which is now a *function*, and
- $F(r) = 1$.

The Feynman-Kac formula enables one to write the solution of (223) as an expectation :

Theorem 11.3. *(Feynman-Kac): Let $V = V(x,t)$ satisfy (223) and (224). Then :*

$$V(\hat{X}_t, t) = \mathbb{E}\left(\exp\left(-\int_t^T c(\hat{X}_u, u)\ du\right) F(\hat{X}_T)|\mathcal{F}_t^W\right), \tag{227}$$

where $\hat{X}_t$ is a solution of the Stochastic Differential Equation :

$$d\hat{X}_t = a(\hat{X}_t, t)dt + \sigma(\hat{X}_t, t)dW_t,\ t \geq 0. \tag{228}$$

Proof. Let $(\hat{X}_t)_t$ satisfy (228). We compute

$$d_u V(\hat{X}_u, u).$$

By Ito's lemma, this equals:

$$\frac{\partial V}{\partial u}((\hat{X}_u, u)du + \frac{\partial V}{\partial x}((\hat{X}_u, u)d\hat{X}_u + \frac{1}{2}\frac{\partial^2 V}{\partial x^2}((\hat{X}_u, u)(d\hat{X}_u)^2.$$

Using (228) and Ito's multiplication rules :

$$(dW_u)^2 = du,\ dW_u du = (du)^2 = 0,$$

we easily find that :

$$d_u V(\hat{X}_u, u) = \left(\frac{\partial V}{\partial u} + \frac{\sigma(\hat{X}_u, u)^2}{2}\frac{\partial^2 V}{\partial x^2} + a(\hat{X}_u, u)\frac{\partial V}{\partial x}\right) du + \sigma(\hat{X}_u, u)dW_u,$$

all derivatives of V evaluated in $(\hat{X}_u, u)$. Since, by hypothesis, V satisfies the PDE (223), this equals :

$$d_u V(\hat{X}_u, u) = c((\hat{X}_u, u)V((\hat{X}_u, u)du + \sigma(\hat{X}_u, u)dW_u.$$

To get rid of the first term on the right hand side, we consider the auxiliary function :

$$w(\hat{X}_t, t) := V(\hat{X}_t, t)e^{-\int_0^t c((\hat{X}_u, u)du}.$$

Then

$$\begin{aligned} d_t w(\hat{X}_t, t) &= \left(d_t V(\hat{X}_t, t) - c(\hat{X}_t, t)\right) e^{-\int_0^t c((\hat{X}_u, u)du} \\ &= f_t dW_t, \end{aligned}$$

where

$$f_t = \sigma(\hat{X}_t, t)e^{-\int_0^t c((\hat{X}_u, u)du}.$$

Integrating from t to T, we find that :

$$w(\hat{X}_T, T) - w(\hat{X}_t, t) = \int_t^T f_u dW_u.$$

We will now take conditional expectations with respect to the natural Brownian filtration. To simplify the notations, we will simply write

$$\mathbb{E}_t(X) \text{ for } \mathbb{E}(X|\mathcal{F}_t^W),$$

although this is somewhat bad notation. Now the key observation is, that the right hand side has conditional expectation $\mathbb{E}_t$ equal to 0:

$$\begin{array}{c}\mathbb{E}_t\left(\int_t^T f_u dW_u\right) = \\ \int_t^T \mathbb{E}_t\left(f_u dW_u\right) = \\ \int_t^T \mathbb{E}_t\left(\mathbb{E}_u\left(f_u dW_u\right)\right) = 0\end{array}$$

since $\mathbb{E}_u(f_u dW_u) = f_u \mathbb{E}_u(dW_u) = 0$: f_u is known at time u, and can therefore be pulled outside the expectation, and $dW_u \sim N(0, du)$ has mean 0 an is independent of $\mathcal{F}_u^W$. Observe that we are simply repeating the proof given on slide VII that Ito-integrals are martingales.

The conclusion therefore is, that

$$\mathbb{E}_t\left(w(\hat{X}_T, T)\right) = w(\hat{X}_t, t). \tag{229}$$

This is basically the Feynman-Kac formula, except that we have to translate things back from w to V. Remembering the definition of w, we see that, from (229,

$$V(\hat{X}_T, T)e^{-\int_0^T c((\hat{X}_u, u)du} = F(\hat{X}_T)e^{-\int_0^T c((\hat{X}_u, u)du},$$

and (229) translates into :

$$V(\hat{X}_t, t)e^{-\int_0^t c((\hat{X}_u, u)du} = \mathbb{E}_t\left(F(\hat{X}_T)e^{-\int_0^T c((\hat{X}_u, u)du}\right).$$

Moving the exponential on the left hand side to the right hand side, and then under the conditional expectation $\mathbb{E}_t$ (which is allowed, since it is known at t), we find formula (223), as required. $$

There is an alternative form of the Feynman-Kac formula, which is convenient for computations, since it does away with the conditional expectation in (223):

Corollary 11.4. *With the notations of theorem 11.3, we have that*

$$V(x,t) = \mathbb{E}\left(\exp\left(-\int_t^T c(\hat{X}_u^x, u)\ du\right) F(\hat{X}_T^x)\right),$$

where $\hat{X}_u^x$ is the solution to the SDE initial value problem:

$$\begin{cases} d\hat{X}_u^x = a(\hat{X}_u, u)du + \sigma(\hat{X}_u^x, u)dW_u,\ u \geq t, \\ \hat{X}_t^x = x \end{cases}$$

The upper index x in $\hat{X}_u^x$ is there to remind you that the solution depends on the initial value x at t.