Stochastic Calculus, Filtering, and Stochastic Control

Lecture Notes

(This version: May 29, 2007)

Ramon van Handel

Spring 2007

Preface

These lecture notes were written for the course ACM 217: *Advanced Topics in Stochastic Analysis* at Caltech; this year (2007), the topic of this course was stochastic calculus and stochastic control in continuous time. As this is an introductory course on the subject, and as there are only so many weeks in a term, we will only consider stochastic integration with respect to the Wiener process. This is sufficient do develop a large class of interesting models, and to develop some stochastic control and filtering theory in the most basic setting. Stochastic integration with respect to general semimartingales, and many other fascinating (and useful) topics, are left for a more advanced course. Similarly, the stochastic control portion of these notes concentrates on verification theorems, rather than the more technical existence and uniqueness questions. I hope, however, that the interested reader will be encouraged to probe a little deeper and ultimately to move on to one of several advanced textbooks.

I have no illusions about the state of these notes—they were written rather quickly, sometimes at the rate of a chapter a week. I have no doubt that many errors remain in the text; at the very least many of the proofs are extremely compact, and should be made a little clearer as is befitting of a pedagogical (?) treatment. If I have another opportunity to teach such a course, I will go over the notes again in detail and attempt the necessary modifications. For the time being, however, the notes are available as-is.

If you have any comments at all about these notes—questions, suggestions, omissions, general comments, and particularly mistakes—I would love to hear from you. I can be contacted by e-mail at `ramon@its.caltech.edu`.

Required background. I assume that the reader has had a basic course in probability theory at the level of, say, Grimmett and Stirzaker [GS01] or higher (ACM 116/216 should be sufficient). Some elementary background in analysis is very helpful.

Layout. The LaTeX layout was a bit of an experiment, but appears to have been positively received. The document is typeset using the `memoir` package and the `daleif1` chapter style, both of which are freely available on the web.

Contents

Introduction

This course is about stochastic calculus and some of its applications. As the name suggests, stochastic calculus provides a mathematical foundation for the treatment of equations that involve noise. The various problems which we will be dealing with, both mathematical and practical, are perhaps best illustrated by considering some simple applications in science and engineering. As we progress through the course, we will tackle these and other examples using our newly developed tools.

Brownian motion, tracking, and finance

Brownian motion and the Wiener process

In 1827, the (then already) famous Scottish botanist Robert Brown observed a rather curious phenomenon [Bro28]. Brown was interested in the tiny particles found inside grains of pollen, which he studied by suspending them in water and observing them under his microscope. Remarkably enough, it appeared that the particles were constantly jittering around in the fluid. At first Brown thought that the particles were alive, but he was able to rule out this hypothesis after he observed the same phenomenon when using glass powder, and a large number of other inorganic substances, instead of the pollen particles. A satisfactory explanation of Brown's observation was not provided until the publication of Einstein's famous 1905 paper [Ein05].

Einstein's argument relies on the fact that the fluid, in which the pollen particles are suspended, consists of a gigantic number of water molecules (though this is now undisputed, the atomic hypothesis was highly controversial at the time). As the fluid is at a finite temperature, kinetic theory suggests that the velocity of every water molecule is randomly distributed with zero mean value (the latter must be the case, as the total fluid has no net velocity) and is independent from the velocity of the other water molecules. If we place a pollen particle in the fluid, then in every time interval the particle will be bombarded by a large number of water molecules, giving it a net random displacement. The resulting random walk of the particle in the fluid is precisely what Brown observed under his microscope.

How should we go about modelling this phenomenon? The following procedure, which is a somewhat modernized version of Einstein's argument, is physically crude but nonetheless quite effective. Suppose that the pollen particle is bombarded by N water molecules per unit time, and that every water molecule contributes an independent, identically distributed (i.i.d.) random displacement ξ_n to the particle (where ξ_n

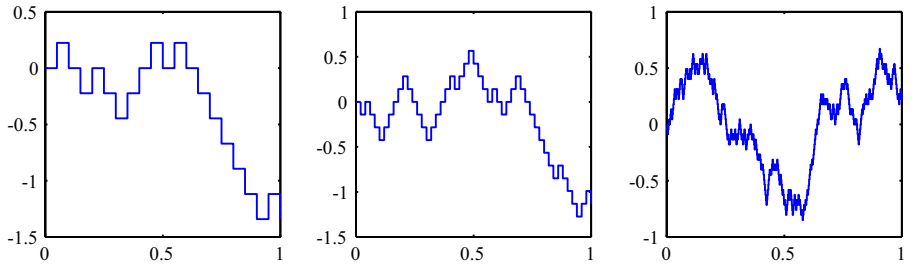

Figure 0.1. Randomly generated sample paths $x_t(N)$ for the Brownian motion model in the text, with (from left to right) $N = 20, 50, 500$ collisions per unit time. The displacements ξ_n are chosen to be random variables which take the values $\pm N^{-1/2}$ with equal probability.

has zero mean). Then at time t, the position $x_t(N)$ of the pollen particle is given by

$$x_t(N) = x_0 + \sum_{n=1}^{\lfloor Nt \rfloor} \xi_n.$$

We want to consider the limit where the number of bombardments N is very large, but where every individual water molecule only contributes a tiny displacement to the pollen particle—this is a reasonable assumption, as the pollen particle, while being small, is extremely large compared to a single water molecule. To be concrete, let us define a constant γ by $\operatorname{var}(\xi_n) = \gamma N^{-1}$. Note that γ is precisely the mean-square displacement of the pollen particle per unit time:

$$\mathbb{E}(x_1(N) - x_0)^2 = \operatorname{var}\left(\sum_{n=1}^{N} \xi_n\right) = N \operatorname{var}(\xi_n) = \gamma.$$

The physical regime in which we are interested now corresponds to the limit $N \to \infty$, i.e., where the number of collisions N is large but the mean-square displacement per unit time γ remains fixed. Writing suggestively

$$x_t(N) = x_0 + \sqrt{\gamma t}\, \frac{\sum_{n=1}^{\lfloor Nt \rfloor} \Xi_n}{\sqrt{Nt}},$$

where $\Xi_n = \xi_n \sqrt{N/\gamma}$ are i.i.d. random variables with zero mean and unit variance, we see that the limiting behavior of $x_t(N)$ as $N \to \infty$ is described by the central limit theorem: we find that the law of $x_t(N)$ converges to a Gaussian distribution with zero mean and variance γt. This is indeed the result of Einstein's analysis.

The limiting motion of the pollen particle as $N \to \infty$ is known as *Brownian motion*. You can get some idea of what $x_t(N)$ looks like for increasingly large N by having a look at figure 0.1. But now we come to our first significant mathematical problem: does the limit of the *stochastic process* $t \mapsto x_t(N)$ as $N \to \infty$ even exist in a suitable sense? This is not at all obvious (we have only shown convergence in

distribution for fixed time t), nor is the resolution of this problem entirely straightforward. If we can make no sense of this limit, there would be no mathematical model of Brownian motion (as we have defined it); and in this case, these lecture notes would come to an end right about here. Fortunately we will be able to make mathematical sense of Brownian motion (chapter 3), which was first done in the fundamental work of Norbert Wiener [Wie23]. The limiting stochastic process x_t (with $\gamma = 1$) is known as the *Wiener process*, and plays a fundamental role in the remainder of these notes.

Tracking a diffusing particle

Using only the notion of a Wiener process, we can already formulate one of the simplest stochastic control problems. Suppose that we, like Robert Brown, are trying to study pollen particles. In order to study the particles in detail, we would like to zoom in on one of the particles—i.e., we would like to increase the magnification of the microscope until one pollen particle fills a large part of the field of view. When we do this, however, the Brownian motion becomes a bit of a nuisance; the random motion of the pollen particle causes it to rapidly leave our field of view. If we want to keep looking at the pollen particle for a reasonable amount of time, we have to keep moving around the cover slide in order to track the motion of the particle.

To deal with this problem, we attach an electric motor to the microscope slide which allows us to move the slide around. Let us call the position of the slide relative to the focus of the microscope z_t; then we can write

$$\frac{dz_t}{dt} = \alpha u_t,$$

where u_t is the voltage applied to the motor and $\alpha > 0$ is a gain constant. The position of the pollen particle relative to the slide is modelled by a Wiener process x_t, so that the position of the particle relative to the microscope focus is given by $x_t + z_t$. We would like to control the slide position to keep the particle in focus, i.e., it is our goal to choose u_t in order that $x_t + z_t$ stays close to zero. To formalize this problem, we could introduce the following cost functional:

$$J_T[u] = p\,\mathbb{E}\left[\frac{1}{T}\int_0^T (x_t + z_t)^2\,dt\right] + q\,\mathbb{E}\left[\frac{1}{T}\int_0^T u_t^2\,dt\right],$$

where p and q are some positive constants. The first term in this expression is the time-average (on some time interval $[0, T]$) mean square distance of the particle from the focus of the microscope: clearly we would like this to be small. The second term, on the other hand, is the average power in the control signal, which should also not be too large in any realistic application (our electric motor will only take so much). The goal of the *optimal control problem* is to find the feedback strategy u_t which minimizes the cost $J_T[u]$. Many variations on this problem are possible; for example, if we are not interested in a particular time horizon $[0, T]$, we could try to minimize

$$J_\infty[u] = p\,\limsup_{T\to\infty}\mathbb{E}\left[\frac{1}{T}\int_0^T (x_t + z_t)^2\,dt\right] + q\,\limsup_{T\to\infty}\mathbb{E}\left[\frac{1}{T}\int_0^T u_t^2\,dt\right].$$

The tradeoff between the conflicting goals of minimizing the distance of the particle to the focus of the microscope and minimizing the feedback power can be selected by modifying the constants p, q. The optimal control theory further allows us to study this tradeoff explicitly: for example, we can calculate the quantity

$$C(U) = \inf \left\{ \limsup_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\int_0^T (x_t + z_t)^2\, dt\right] : \limsup_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\int_0^T u_t^2\, dt\right] \le U \right\},$$

i.e., $C(U)$ is the smallest time-average tracking error that is achievable using controls whose time-average power is at most U. This gives a fundamental limit on the performance of our tracking system under power constraints. The solution of these problems, in a much more general context, is the topic of chapter 6.

The stock market: how to invest your money

Though the theoretical ideas behind Brownian motion are often attributed to Einstein, the same model was developed several years earlier in a completely different context by the French mathematician Louis Bachelier [Bac00].[1] Bachelier was interested in speculation on *rentes* (French government bonds), and introduced the Brownian motion to model the fluctuations in the bond prices. Bachelier's work forms the foundation for much of the modern theory of mathematical finance, though his work was virtually unknown to economists for more than half a century. These days mathematical finance is an important application area for stochastic analysis and stochastic control, and provides a rich source of interesting problems.

We will only consider stock. Companies issue stock in order to finance their operations; the money made from the sale of stock can then be used by the company to finance production, special projects, etc. In return, a certain portion of the profit made by the company is periodically paid out to the shareholders (people who own stock in the company). Such payments are called dividends. If the company is doing well (e.g., if sales are soaring), then owning stock in the company is likely to be profitable.

This is only the beginning of the story, however. Any individual who owns stock in a company can decide to sell his stock on a stock market. As you can imagine, the going rate for a particular stock depends on how well the company is doing (or is expected to do in the future). When the company is doing well, many people would like to own stock (after all, there is a prospect of large dividends) while few people who own stock are willing to sell. This drives up the market price of the stock. When the company is not doing well, however, it is likely that more shareholders are willing to sell than there is demand for the stock, so that the market price of the stock is low. Due to these "market forces", the stock prices tend to fluctuate randomly in the course of time; see, for example, figure 0.2. Even if we ignore dividends (which we will do to simplify matters), we can still try to make money on the stock market by buying stock when the price is low and selling when the price is high.

There are now many interesting and pertinent questions that we can ask. For example, how should we invest our money in the stock market to maximize our profit?

[1] An excellent annotated translation has recently appeared in [DE06].

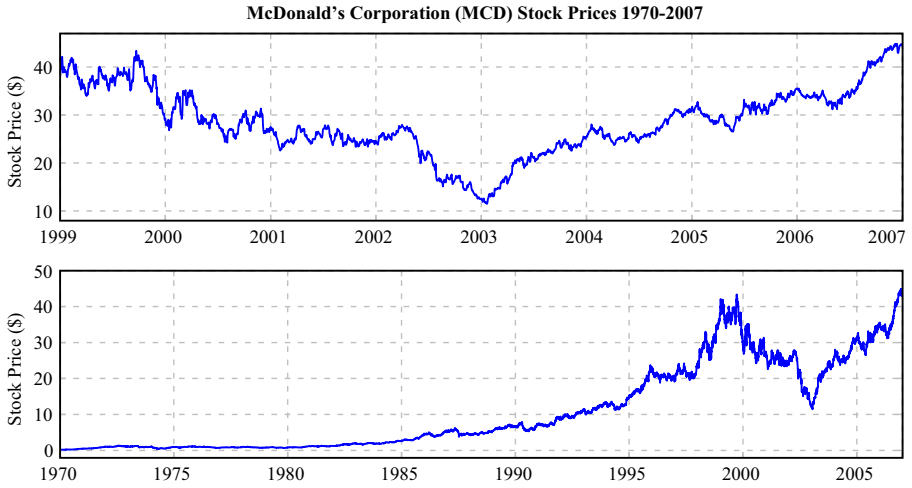

Figure 0.2. Market price of McDonald's stock on the New York Stock Exchange (NYSE) over the period 1999–2007 (upper plot) and 1970–2007 (lower plot). The stock price data for this figure was obtained from *Yahoo! Finance* at `finance.yahoo.com`.

This is essentially a stochastic control problem, which was tackled in a famous paper by Merton [Mer71]. A different class of questions concerns the pricing of options—a sort of "insurance" issued on stock—and similar financial derivatives. The modern theory of option pricing has its origins in the pioneering work of Black and Scholes [BS73] and is an important problem in practice. There are many variations on these and other topics, but they have at least one thing in common: their solution requires a healthy dose of stochastic analysis.

For the time being, let us consider how to build a mathematical model for the stock prices—a first step for further developments. Bachelier used Brownian motion for this purpose. The problem with that model is that the stock prices are not guaranteed to be positive, which is unrealistic; after all, nobody pays money to dispose of his stock. Another issue to take into account is that even though this is not as visible on shorter time scales, stock prices tend to grow exponentially on the long run: see figure 0.2. Often this exponential rate will be larger than the interest rate we can obtain by putting our money in the bank, which is a good reason to invest in stock (investing in stock is not the same as gambling at a casino!) This suggests the following model for stock prices, which is widely used: the price S_t at time t of a single unit of stock is given by

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

where W_t is a Wiener process, and $S_0 > 0$ (the initial price), $\mu > 0$ (the return rate), and $\sigma > 0$ (the volatility) are constants. A stochastic process of this form is called *geometric Brownian motion*. Note that S_t is always positive, and moreover $\mathbb{E}(S_t) = S_0 e^{\mu t}$ (exercise: you should already be able to verify this!) Hence evidently,

on average, the stock makes money at rate μ. In practice, however, the price may fluctuate quite far away from the average price, as determined by magnitude of the volatility σ. This means that there is a probability that we will make money at a rate much faster than μ, but we can also make money slower or even lose money. In essence, a stock with large volatility is a risky investment, whereas a stock with small volatility is a relatively sure investment. When methods of mathematical finance are applied to real-world trading, parameters such as μ and σ are often estimated from real stock market data (like the data shown in figure 0.2).

Beside investing in the stock S_t, we will also suppose that we have the option of putting our money in the bank. The bank offers a fixed interest rate $r > 0$ on our money: that is, if we initially put R_0 dollars in the bank, then at time t we will have

$$R_t = R_0 \exp(rt)$$

dollars in our bank account. Often it will be the case that $r < \mu$; that is, investing in the stock will make us more money, on average, than if we put our money in the bank. On the other hand, investing in stock is risky: there is some finite probability that we will make less money than if we had invested in the bank.

Now that we have a model for the stock prices and for the bank, we need to be able to calculate how much money we make using a particular investment strategy. Suppose that we start initially with a capital of X_0 dollars. We are going to invest some fraction θ_0 of this money in stock, and the rest in the bank (i.e., we put $(1-\theta_0)X_0$ dollars in the bank, and we buy $\theta_0 X_0/S_0$ units of stock). Then at time t, our total wealth X_t (in dollars) amounts to

$$X_t = \theta_0 X_0\, e^{(\mu-\sigma^2/2)t+\sigma W_t} + (1-\theta_0)X_0 e^{rt}.$$

Now suppose that at this time we decide to reinvest our capital; i.e., we now invest a fraction θ_t of our newly accumulated wealth X_t in the stock (we might need to either buy or sell some of our stock to ensure this), and put the remainder in the bank. Then at some time $t' > t$, our total wealth becomes

$$X_{t'} = \theta_t X_t\, e^{(\mu-\sigma^2/2)(t'-t)+\sigma(W_{t'}-W_t)} + (1-\theta_t)X_t e^{r(t'-t)}.$$

Similarly, if we choose to reinvest at the times $0 = t_0 < t_1 < \cdots < t_{N-1} < t_N = t$, then we find that our wealth at time t is given by

$$X_t = X_0 \prod_{n=1}^{N} \left[\theta_{t_{n-1}}\, e^{(\mu-\sigma^2/2)(t_n-t_{n-1})+\sigma(W_{t_n}-W_{t_{n-1}})} + (1-\theta_{t_{n-1}})e^{r(t_n-t_{n-1})}\right].$$

In principle, however, we should be able to decide at any point in time what fraction of our money to invest in stock; i.e., to allow for the most general trading strategies we need to generalize these expressions to the case where θ_t can vary continuously in time. At this point we are not equipped to do this: we are missing a key mathematical ingredient, the *stochastic calculus* (chapters 4 and 5). Once we have developed the latter, we can start to pursue the answers to some of our basic questions.

White noise, corrupted signals, and noise-driven systems

White noise and the Wiener process

Despite the fact that its mathematical description is somewhat elusive, the notion of white noise is used widely in science and engineering. Indeed, you have most likely encountered this idea in some form or another in the past. We will revisit it now and exhibit some of the associated difficulties.

Perhaps the simplest notion of white noise is the one used in discrete time. Suppose that we have some discrete time message $\{a_n\}$ which we would like to transmit to a receiver. During the transmission, however, the message becomes corrupted: for example, any signal transmitted through a wire is subject to thermal noise, whereas signals sent through radio-frequency transmission are subject to all sorts of environmental disturbances. Now it is very often the case that each letter a_n of the message is essentially corrupted independently. For example, the atmospheric disturbances tend to fluctuate much faster than the rate at which we transmit the letters in our message, so by the time we transmit the next letter a_{n+1} we see a completely different disturbance. In this case, we would model the signal observed by the receiver by $x_n = a_n + \xi_n$, where $\{\xi_n\}$ are i.i.d. random variables with zero mean. If, in addition, we assume that every disturbance is itself generated by many independent small effects, then the central limit theorem suggests that ξ_n should be Gaussian random variables. In this case, we say that $\{\xi_n\}$ is *discrete time white noise*, or AWGN ("additive white Gaussian noise") in the language of communications theory.

How should we generalize this to the continuous time case? A first idea would be to attach to every time $t \in \mathbb{R}_+$ an i.i.d. zero mean Gaussian random variable ξ_t (with unit variance, say), just like we did in the discrete time case. Evidently we would have $\mathbb{E}(\xi_s \xi_t) = 0$ if $s \neq t$ and $\mathbb{E}(\xi_t^2) = 1$. Even putting aside the issue of mathematical well-posedness of this process, we can see immediately that it would not be of much use. Let us, hypothetically, take our process ξ_t and pass it through a signal processing device which calculates its time average Ξ_ε over an arbitrarily small interval $[0, \varepsilon]$:

$$\Xi_\varepsilon = \frac{1}{\varepsilon} \int_0^\varepsilon \xi_t \, dt.$$

Then obviously $\mathbb{E}(\Xi_\varepsilon) = 0$, but also

$$\text{var}(\Xi_\varepsilon) = \mathbb{E}(\Xi_\varepsilon^2) = \frac{1}{\varepsilon^2} \int_0^\varepsilon \int_0^\varepsilon \mathbb{E}(\xi_s \xi_t) \, ds \, dt = 0.$$

Hence suppose we transmit a letter a_0 of our message in white noise: $x_t = a_0 + \xi_t$. Then after an arbitrarily small time ε (i.e. as soon as we have data, however little), we would be able to know exactly what a_0 was simply by calculating the time average of x_t. Clearly ξ_t does not qualify as *noise*, so we dump it in the stack of bad ideas.[2]

[2] Mathematically, the process ξ_t suggested in this paragraph is a mess. One could, at least in principle, construct such a process using a technique known as Kolmogorov's extension theorem. However, it turns out that there is no way to do this in such a way that the sample paths $t \mapsto \xi_t$ are even remotely well-behaved: such paths can never be *measurable* [Kal80, Example 1.2.5]. In particular, this implies that there is, even in principle, no way in which we could possibly make mathematical sense of the time average of this process (integration requires measurability, the integral of a non-measurable function is meaningless). This resolves our little paradox, but also highlights that this sort of construction is manifestly useless.

How, then, should we define white noise ξ_t in a meaningful way? Having learned our lesson from the previous example, we might try to insist that the time average of ξ_t is well defined. Inspired by AWGN, where in unit time the corrupting noise is a zero mean Gaussian random variable (with unit variance, say), we could require that the average white noise in unit time Ξ_1 is a zero mean Gaussian random variable with unit variance. We also want to insist that ξ_t retains its independence property: ξ_t and ξ_s are i.i.d. for $t \neq s$. This means, in particular, that

$$\int_0^{1/2} \xi_t \, dt \qquad \text{and} \qquad \int_{1/2}^1 \xi_t \, dt$$

must both be Gaussian random variables with mean zero and variance $1/2$ (after all, their sum equals Ξ_1 and they must be i.i.d.), etc. Proceeding along these lines (convince yourself of this!), it is not difficult to conclude that

$$\int_0^t \xi_s \, ds = W_t \quad \textit{must be a Wiener process.}$$

Hence we conjecture that the correct "definition" of white noise is: ξ_t is the time derivative dW_t/dt of a Wiener process W_t. Unfortunately for us, the Wiener process turns out to be non-differentiable for almost every time t. Though we cannot prove it yet, this is easily made plausible. Recall that W_t is a Gaussian random variable with variance t; to calculate dW_t/dt at $t = 0$, for example, we consider W_t/t and let $t \to 0$. But W_t/t is a Gaussian random variable with variance t^{-1}, so clearly something diverges as $t \to 0$. Apparently, we are as stuck as before.

Let us explore a little bit further. First, note that the covariance of the Wiener process is[3] $\mathbb{E}(W_s W_t) = s \wedge t$. To see this, it suffices to note that for $s \leq t$, $W_t - W_s$ and W_s are independent (why?), so $\mathbb{E}(W_s W_t) = \mathbb{E}(W_s^2) + \mathbb{E}(W_s(W_t - W_s)) = s$. Let us now formally compute the covariance of white noise:

$$\mathbb{E}(\xi_s \xi_t) = \frac{d}{dt}\frac{d}{ds}\mathbb{E}(W_s W_t) = \frac{d}{dt}\frac{d}{ds}\frac{s+t-|t-s|}{2} = \frac{d}{dt}\frac{1+\text{sign}(t-s)}{2} = \delta(t-s),$$

where $\delta(t)$ is the Dirac delta "function". This is precisely the defining property of white noise as it is used in the engineering literature and in physics. Of course, the non-differentiability of the Wiener process is driven home to us again: as you well know, the Dirac delta "function" is not actually a function, but a distribution (generalized function), an object that we could never get directly from the theory of stochastic processes. So the bad news is that despite the widespread use of white noise,

a mathematical model for white noise *does not exist*,

at least within the theory of stochastic processes.[4] Fortunately there is also some good news: as we will see below and throughout this course,

[3] The lattice-theoretic notation $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$ is very common in the probability literature. We will adopt it throughout this course.

[4] We could generalize our notion of a stochastic process to include random objects whose "sample paths" can be, for example, tempered distributions. In the class of generalized stochastic processes one can make sense of white noise (see [Hid80, HØUZ96], or [Arn74, sec. 3.2] for a simple introduction). This is not particularly helpful, however, in most applications, and we will not take this route.

> almost anything we would like to do with white noise, including its applications in science and engineering, can be made rigorous by working directly with the Wiener process.

For example, consider the transmission of a continuous-time signal a_t through white noise ξ_t: i.e., the corrupted signal is formally given by $x_t = a_t + \xi_t$. This quantity is not mathematically meaningful, but we can integrate both sides and obtain

$$X_t = \int_0^t x_s \, ds = \int_0^t a_s \, ds + \int_0^t \xi_s \, ds = \int_0^t a_s \, ds + W_t.$$

The right-hand side of this expression is mathematically meaningful and does not involve the notion of white noise. At least formally, the process X_t should contain the same information as x_t: after all, the latter is obtained from the former by formal differentiation. If we want to estimate the signal a_t from the observations x_t, we might as well solve the same problem using X_t instead—the difference being that the latter is a mathematically well-posed problem.

Why do we insist on using white noise? Just like in mathematics, true white noise does not exist in nature; any noise encountered in real life has fairly regular sample paths, and as such has some residual correlations between the value of the process at different times. In the majority of applications, however, the correlation time of the noise is very short compared to the other time scales in the problem: for example, the thermal fluctuations in an electric wire are much faster than the rate at which we send data through the wire. Similar intuition holds when we consider a dynamical system, described by a differential equation, which is driven by noise whose random fluctuations are much faster than the characteristic timescales of the dynamics (we will return to this below). In such situations, the idea that we can approximate the noise by white noise is an extremely useful idealization, even if it requires us to scramble a little to make the resulting models fit into a firm mathematical framework.

The fact that white noise is (formally) independent at different times has far-reaching consequences; for example, dynamical systems driven by white noise have the Markov property, which is not the case if we use noise with a finite correlation time. Such properties put extremely powerful mathematical tools at our disposal, and allow us to solve problems in the white noise framework which would be completely intractable in models where the noise has residual correlations. This will become increasingly evident as we develop and apply the necessary mathematical machinery.

Tracking revisited

Let us return for a moment to the problem of tracking a diffusing particle through a microscope. Previously we tried to keep the particle in the field of view by modifying the position of the microscope slide based on our knowledge of the location of the particle. In the choice of a feedback strategy, there was a tradeoff between the necessary feedback power and the resulting tracking error.

The following is an interesting variation on this problem. In biophysics, it is of significant interest to study the dynamical properties of individual biomolecules—such as single proteins, strands of DNA or RNA—in solution. The dynamical properties

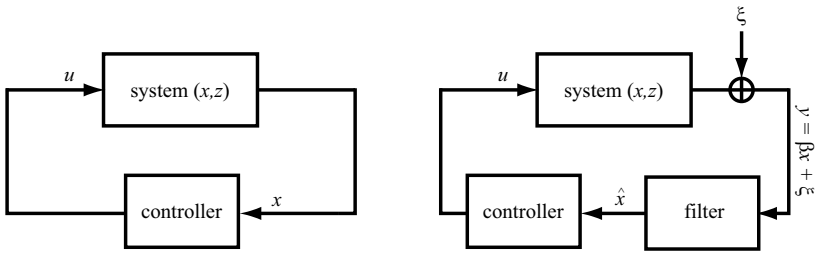

Figure 0.3. Optimal feedback control strategies with full (left) and partial (right) information. In the full information case, the control signal is a function of the state of the system. In the partial information case, the observation process is first used to form an estimate of the state of the system; the control signal is then a function of this estimate.

of these molecules provide some amount of insight into protein folding, DNA replication, etc. Usually, what one does is to attach one or several fluorescent dyes to the molecules of interest. Using a suitably designed microscope, a laser beam is focused on a dilute sample containing such molecules, and the fluorescent light is captured and detected using a photodetector. The molecules perform Brownian motion in the solution, and occasionally one of the molecules will randomly drift into the focus of the laser beam. During the period of time that the molecule spends in the focus of the beam, data can be collected which is subsequently analyzed to search for signatures of the dynamical behavior of interest. The problem is that the molecules never stay in the beam focus very long, so that not much data can be taken from any single molecule at a time. One solution to this problem is to anchor the molecules to a surface so that they cannot move around in solution. It is unclear, however, that this does not significantly modify the dynamical properties of interest.

A different solution—you guessed it—is to try to follow the molecules around in the solution by moving around the microscope slide (see [BM04] and references therein). Compared to our previous discussion of this problem, however, there is now an additional complication. Previously we assumed that we could see the position of the particle under the microscope; this information determined how we should choose the control signal. When we track a single molecule, however, we do not really "see" the molecule; the only thing available to us is the fluorescence data from the laser, which is inherently noisy ("shot noise"). Using a suitable modulation scheme [BM04], we can engineer the system so that to good approximation the position data at our disposal is given by $y_t = \beta x_t + \xi_t$, where ξ_t is white noise, x_t is the distance of the molecule to the center of the slide, and β is the signal-to-noise ratio. As usual, we make this rigorous by considering the integrated observation signal

$$Y_t = \int_0^t y_s \, ds = \int_0^t \beta x_s \, ds + W_t.$$

Our goal is now to minimize a cost function of the form $J_\infty[u]$, for example, but with an additional constraint: our feedback strategy u_t is only allowed to depend on the

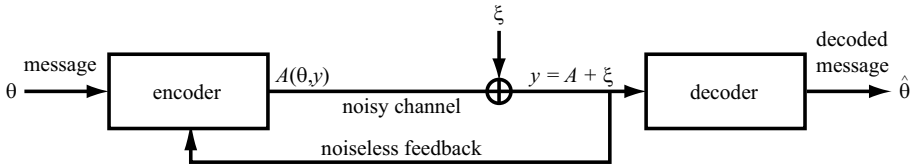

Figure 0.4. Setup for the transmission of a message over a noisy channel. We get to choose the encoder and the decoder; moreover, the encoder has access to the corrupted signal observed by the decoder. How should we design the system to minimize the transmission error?

past observations, i.e., on the process Y_s on the interval $s \in [0, t]$. Compared to our previous discussion, where we were allowed to base our feedback strategy directly on the particle position x_s, we have apparently lost information, and this will additionally limit the performance of our control strategy.

At first sight, one could expect that the optimal control strategy in the case of observation-based feedback would be some complicated functional of the observation history. Fortunately, it turns out that the optimal control has a very intuitive structure, see figure 0.3. The controller splits naturally into two parts. First, the observations are used to form an estimate of the state of the system (i.e., the position of the molecule). Then the control signal is chosen as a function of this estimate. This structure is quite universal, and is often referred to as the *separation principle* of stochastic control. It also highlights the fact that *filtering*—the estimation of a stochastic process from noisy observations—is intimately related with stochastic control. Filtering theory is an interesting and important topic on its own right; it will be studied in detail in chapter 7, as well as the connection with control with partial observations.

Transmitting a message over a noisy channel

A somewhat nonstandard control problem appears in communications theory, see figure 0.4. We would like to transmit a message θ_t over a noisy channel. If we were to send the message directly over the channel, the receiver would see the corrupted message $y_t = \theta_t + \xi_t$ where ξ_t is white noise; what remains is a filtering problem, where the receiver attempts to form an estimate $\hat{\theta}_t$ of the message from the noisy signal.

We are free, however, to encode the message in any way we want at the transmitter side: i.e., we transmit the encoded message $A_t(\theta)$ over the channel, so that the receiver sees $y_t = A_t(\theta) + \xi_t$. We then have to design a decoder on the other end to extract the encoded message appropriately from the noisy signal. The problem becomes even more interesting if the encoder can see the corrupted signal observed by the receiver; in this case, the encoded message takes the form $A_t(\theta, y)$ (we speak of a noisy channel with noiseless feedback). The questions are obvious: how should we design the encoder-decoder pair to minimize the error between the true message θ_t and the decoded message $\hat{\theta}_t$, and what is the optimal performance? Note that just as in the tracking example, we will have to impose some restriction on the signal power,

for example, we could impose a constraint of the form

$$\mathbb{E}\left[\frac{1}{T}\int_0^T A_t(\theta, y)^2\, dt\right] \leq U.$$

After all, if we do not do this then we could transmit an arbitrarily strong signal through the channel, and an optimal solution would be to transmit the message θ_t directly through the channel with an infinite signal-to-noise ratio. With a power constraint in place, however, there will be a fundamental limitation on the achievable performance. We will see that these problems can be worked out in detail, for example, if the message θ_t is modelled as a Gaussian process (see, e.g., [LS01b]).

Changepoint detection and hypothesis testing

Suppose we have a device which exhibits an abrupt change of behavior at a certain random time. For example, imagine a production line in a factory where one of the machines along the line breaks down; a computer network that suddenly gets attacked by hackers; or an atom in a quantum optics laboratory which spontaneously emits a photon. We would like to detect when the change occurs so that we can take appropriate measures; for example, if one of the machines in the factory breaks down, we should fix it. Often, however, we cannot directly observe when the change occurs; all we have available to us is noisy data. In the factory case, we would observe what percentage of the production is defective; in the network case we are trying to observe a change in the network traffic; and in the case of an atom we are observing a photocurrent with the associated shot noise. In all these cases we are faced with the problem of distinguishing an abnormal change from the normal fluctuations in the observed data. In statistics this is known as the problem of *changepoint detection*.

To model the problem, suppose that the abrupt change occurs at a time τ, which is randomly distributed according to a suitable law. The signal that we are allowed to observe is of the form $y_t = \alpha \mathbf{1}_{t\geq\tau} + \xi_t$, where $\alpha > 0$, ξ_t is white noise and $\mathbf{1}_{t\geq\tau} = 1$ if $t \geq \tau$ and 0 otherwise. Our goal is to find a time ϑ which depends only on the observations y_t and which is close to τ in an appropriate sense.

In choosing ϑ there are two conflicting considerations. First, we would like to minimize the probability of $\vartheta < \tau$, i.e., of deciding to intervene before the change has actually occured. Clearly trying to repair a machine which is operating perfectly well is a waste of time and money. On the other hand, if $\vartheta \geq \tau$, we do not want the detection delay $\vartheta - \tau$ to be too large; if we wait longer than necessary to repair the machine, we will waste expensive material and produce a lot of defective merchandise. To formalize these considerations, we could introduce a cost functional of the form

$$J[\vartheta] = p\,\mathbb{P}[\vartheta < \tau] + q\,\mathbb{E}[\vartheta - \tau | \vartheta \geq \tau]\,\mathbb{P}[\vartheta \geq \tau].$$

The goal is then to choose ϑ that minimizes $J[\vartheta]$, and the choice of p and q determine the relative importance of achieving a low false alarm rate or a short detection delay. Alternatively, we could ask: given that we tolerate a fixed false alarm rate, how should we choose ϑ is minimize the detection delay? Note that these considerations are very similar to the ones we discussed in the tracking problem, where the tradeoff was

between achieving good tracking performance and low feedback power. Indeed, in many ways this type of problem is just like a control problem, except that the control action in this case is the time at which we decide to intervene rather than the choice of a feedback strategy. Similarly, the separation principle and filtering theory play an important role in the solution of this problem (due to Shiryaev [Shi73]).

A related idea is the problem of *hypothesis testing*. Here we are given noisy data, and our job is to decide whether there is a signal buried in the noise. To be more precise, we have two hypotheses: under the null hypothesis H_0, the observations have the form $y_t = \xi_t$ where ξ_t is white noise; under the alternative hypothesis H_1 there is a given signal θ_t buried in the noise, i.e., $y_t = \theta_t + \xi_t$. Such a scenario occurs quite often, for example, in surveillance and target detection, gravitational wave detection, etc. It is our goal to determine whether the hypothesis H_0 or H_1 is correct after observing y_t some time ϑ. Once again we have conflicting interests; on the one hand, we would like to make a pronouncement on the matter as soon as possible (ϑ should be small), while on the other hand we would like to minimize the probability that we obtain the wrong answer (we choose H_0 while H_1 is true, and vice versa). The rest of the story is much as before; for example, we can try to find the hypothesis test which takes the least time ϑ under the constraint that we are willing to tolerate at most some given probability α of selecting the wrong hypothesis.

Both these problems are examples of *optimal stopping* problems: control problems where the goal is to select a suitable stopping time ϑ. Optimal stopping theory has important applications not only in statistics, but also in mathematical finance and in several other fields. We can also combine these ideas with more traditional control theory as follows. Suppose that we wish to control a system (for example our favorite tracking system) not by applying continuous feedback, but by applying feedback impulses at a set of discrete times. The question now becomes: at which times can we best apply the control, and what control should we apply at those times? Such problems are known as *impulse control* problems, and are closely related to optimal stopping problems. Optimal stopping and impulse control are the topics of chapter 8.

Stochastic differential equations

In the previous sections we have discussed some applications of the Wiener process and its use in white noise modelling. In each example, the Brownian motion or white noise were used "as is"; in the tracking and finance examples the dynamics of interest was itself a Wiener process, whereas in the remaining examples pure white noise was used as a model for signal corruption. We have left for last what is perhaps the most important (and widely used) form of stochastic modelling in continuous time: the use of white noise as a driving force for differential equations. The theory of stochastic differential equations is extremely flexible and emerges naturally in a wide range of applications; it is thus not surprising that it is the basic tool in the modelling and analysis of a large number of stochastic systems.

The basic idea is very simple: we would like to give meaning to the solution x_t of

$$\frac{dx_t}{dt} = b(t, x_t) + \sigma(t, x_t)\,\xi_t,$$

where b and σ are given (sufficiently smooth) functions and ξ_t is white noise. Such an equation could be interesting for many reasons; in particular, one would expect to obtain such an equation from any deterministic model

$$\frac{dx_t}{dt} = b(t, x_t) + \sigma(t, x_t)\, u_t,$$

where u_t is some input to the system, when the input signal is noisy (but we will see that there are some subtleties here—see below). If $\sigma = 0$, our stochastic equation is just an ordinary differential equation, and we can establish existence and uniqueness of solutions using standard methods (notably Picard iteration for the existence question). When $\sigma \neq 0$, however, the equation as written does not even make sense: that infernal nuisance, the formal white noise ξ_t, requires proper interpretation.

Let us first consider the case where $\sigma(t, x) = \sigma$ is a constant, i.e.,

$$\frac{dx_t}{dt} = b(t, x_t) + \sigma\xi_t.$$

This *additive noise* model is quite common: for example, if we model a particle with mass m in a potential $V(x)$, experiencing a noisy force $F_t = \sigma\xi_t$ (e.g., thermal noise) and a friction coefficient k, then the particle's position and momentum satisfy

$$\frac{dx_t}{dt} = m^{-1}p_t, \qquad \frac{dp_t}{dt} = -\frac{dV}{dx}(x_t) - kp_t + \sigma\xi_t.$$

How should we interpret such an equation? Let us begin by integrating both sides:

$$x_t = x_0 + \int_0^t b(s, x_s)\, ds + \sigma W_t,$$

where W_t is a Wiener process. Now this equation makes sense! We will say that the stochastic differential equation

$$dx_t = b(t, x_t)\, dt + \sigma\, dW_t$$

has a unique solution, if there is a unique stochastic process x_t that satisfies the associated integral equation. The differential notation dx_t, etc., reminds us that we think of this equation as a sort of differential equation, but it is important to realize that this is just notation: stochastic differential equations are not actually differential equations, but integral equations like the one above. We could never have a "real" stochastic differential equation, because clearly x_t cannot be differentiable!

For additive noise models, it is not difficult to establish existence and uniqueness of solutions; in fact, one can more or less copy the proof in the deterministic case (Picard iteration, etc.) However, even if we can give meaning to such an equation, we are lacking some crucial analysis tools. In the deterministic theory, we have a key tool at our disposal that allows us to manipulate differential equations: undergraduate calculus, and in particular, that wonderful chain rule! Here, however, the chain rule will get us in trouble; for example, let us *naively* calculate the equation for x_t^2:

$$\frac{d}{dt}x_t^2 \stackrel{?}{=} 2x_t\frac{dx_t}{dt} = 2x_t\, b(t, x_t) + 2\sigma x_t\, \xi_t.$$

This is no longer an additive noise model, and we are faced with the problem of giving meaning to the rather puzzling object

$$\int_0^t x_s \, \xi_s \, ds \;=\; ??.$$

The resolution of this issue is key to almost all of the theory in this course! Once we have a satisfactory definition of such an integral (chapter 4), we are in a position to define general stochastic differential equations (chapter 5), and to develop a *stochastic calculus* that allows us to manipulate stochastic differential equations as easily as their deterministic counterparts. Here we are following in the footsteps of Kiyosi Itô [Itô44], whose name we will encounter frequently throughout this course.

In chapter 4 we will define a new type of integral, the *Itô integral*

$$\int_0^t x_s \, dW_s,$$

which will play the role of a white noise integral in our theory. We will see that this integral has many nice properties; e.g., it has zero mean, and will actually turn out to be a martingale. We will also find a change of variables formula for the Itô integral, just like the chain rule in ordinary calculus. The Itô change of variables formula, however, is not the same as the ordinary chain rule: for example, for any $f \in C^2$

$$df(W_t) = f'(W_t)\, dW_t + \tfrac{1}{2} f''(W_t)\, dt,$$

while the usual chain rule would only give the first term on the right. This is not surprising, however, because the ordinary chain rule cannot be correct (at least if we insist that our stochastic integral has zero mean). After all, if the chain rule were correct, the variance of W_t could be calculated as

$$\mathrm{var}(W_t) = \mathbb{E}(W_t^2) = \mathbb{E}\left[2 \int_0^t W_s \, dW_s\right] = 0,$$

which is clearly untrue. The formula above, however, gives the correct answer

$$\mathrm{var}(W_t) = \mathbb{E}(W_t^2) = \mathbb{E}\left[2 \int_0^t W_s \, dW_s + \int_0^t ds\right] = t.$$

Evidently things work a little differently in the stochastic setting than we are used to; but nonetheless our tools will be almost as powerful and easy to use as their deterministic counterparts—as long as we are careful!

The reader is probably left wondering at this point whether we did not get a little carried away. We started from the intuitive idea of an ordinary differential equation driven by noise. We then concluded that we can not make sense of this as a true differential equation, but only as an integral equation. Next, we concluded that we didn't really know what this integral is supposed to be, so we proceeded to make one up. Now we have finally reduced the notion of a stochastic differential equation to a mathematically meaningful form, but it is unclear that the objects we have introduced bear any resemblance to the intuitive picture of a noisy differential equation.

To justify our models, let us consider a differential equation driven by a random process ξ_t^ε with (piecewise) continuous sample paths:

$$\frac{dx_t^\varepsilon}{dt} = b(t, x_t^\varepsilon) + \sigma(t, x_t^\varepsilon)\, \xi_t^\varepsilon.$$

This is just a nonautonomous ordinary differential equation, and can be solved in the usual way. The idea is now to assume that ξ_t^ε fluctuates so fast, that it is well approximated by white noise: to be more precise, we assume that $\xi_t^\varepsilon = dW_t^\varepsilon/dt$, where W_t^ε converges to a Wiener process in a suitable sense as $\varepsilon \to 0$. Obviously the limit of ξ_t^ε as $\varepsilon \to 0$ cannot exist. Nonetheless the limit of x_t^ε as $\varepsilon \to 0$ is usually perfectly well defined (in a suitable sense), and $x_t = \lim_{\varepsilon\to 0} x_t^\varepsilon$ can in fact be shown to satisfy an Itô stochastic differential equation. Hence our use of the Itô theory is well justified in hindsight: we can indeed use it to approximate differential equations driven by rapidly fluctuating non-white noise. There are significant advantages to making the white noise approximation, however: for one, the process x_t turns out to be a *Markov* process, whereas this is certainly not the case for x_t^ε. The Markov property is crucial in the development of stochastic control and filtering theory—these and many other developments would be completely intractable if we worked directly with x_t^ε.

What is perhaps surprising is that the limiting equation for x_t is not the one we expect. In fact, x_t will satisfy the stochastic differential equation [WZ65]

$$dx_t = b(t, x_t)\, dt + \tfrac{1}{2}\sigma'(t, x_t)\sigma(t, x_t)\, dt + \sigma(t, x_t)\, dW_t,$$

where $\sigma'(t, x) = d\sigma(t, x)/dx$. The second term on the right is known as the Wong-Zakai correction term, and our naive interpretation of stochastic differential equations cannot account for it! Nonetheless it is not so strange that it is there. To convince yourself of this, note that x_t^ε must satisfy the ordinary chain rule: for example,

$$\frac{dx_t^\varepsilon}{dt} = Ax_t^\varepsilon + Bx_t^\varepsilon\, \xi_t^\varepsilon, \qquad \frac{d(x_t^\varepsilon)^2}{dt} = 2A(x_t^\varepsilon)^2 + 2B(x_t^\varepsilon)^2\, \xi_t^\varepsilon.$$

If we take the limit as $\varepsilon \to 0$, we get using the Wong-Zakai correction term

$$dx_t = (A+\tfrac{1}{2}B^2)x_t\, dt + Bx_t\, dW_t, \qquad d(x_t)^2 = (2A+2B^2)(x_t)^2 dt + 2B(x_t)^2 dW_t.$$

If the ordinary chain rule held for x_t as well, then we would be in trouble: the latter equation has an excess term $B^2(x_t)^2 dt$. But the ordinary chain rule does not hold for x_t, and the additional term in the Itô change of variables formula gives precisely the additional term $B^2(x_t)^2 dt$. Some minor miracles may or may not have occured, but at the end of the day everything is consistent—as long as we are sufficiently careful!

Regardless of how we arrive at our stochastic differential equation model—be it through some limiting procedure, through an empirical modelling effort, or by some other means—we can now take such an equation as our starting point and develop stochastic control and filtering machinery in that context. Almost all the examples that we have discussed require us to use stochastic differential equations at some point in the analysis; it is difficult to do anything without these basic tools. If you must choose to retain only one thing from this course, then it is this: remember how stochastic calculus and differential equations work, because they are ubiquitous.

An outline of this course

Now that you have a flavor of things to come, little remains but to dive in. This introductory chapter has necessarily been a little vague at points; things will become increasingly clear and precise as we make our way through the theory.

We will begin, in chapter 1, by reviewing the basic tools of mathematical probability theory. Perhaps the theory will be presented a little more formally than you have seen in previous courses, but a measure-theoretic approach to probability will be indispensible in the remaining chapters. Chapter 2 introduces some more of the basic tools: conditional expectations, martingales, stochastic processes, and stopping times. Chapters 1 and 2 together provide a crash course in the fundamentals of probability theory; much of this material may be known to you already—the more the better!

In chapter 3 we will discuss the Wiener process. Mostly we will prove that it actually exists—a nontrivial exercise!—and investigate some of its properties.

Chapters 4 and 5 are the most important chapters of this course. They introduce the Itô integral and stochastic differential equations, respectively. We will also discuss some of the most important theorems in stochastic analysis, Girsanov's theorem and the martingale representation theorem, that you absolutely cannot live without. On the side we will learn some useful tricks, such as how to simulate stochastic differential equations in MATLAB, and how Lyapunov function methods can be extended to the stochastic case (which allows us to do some simple nonlinear stochastic control).

The remainder of the course centers around stochastic control and filtering. Chapter 6 introduces the basic methods of optimal stochastic control, which will allow us to solve problems such as the tracking example (with full observations) and some problems in finance. Chapter 7 develops filtering theory and its connection with control. Finally, chapter 8 discusses optimal stopping and impulse control problems.

CHAPTER 1

Review of Probability Theory

This chapter is about basic probability theory: probability spaces, random variables, limit theorems. Much of this material will already be known to you from a previous probability course. Nonetheless it will be important to formalize some of the topics that are often treated on a more intuitive level in introductory courses; particularly the measure-theoretic apparatus, which forms the foundation for mathematical probability theory, will be indispensible. If you already know this material, you can skip to chapter 2; if not, this chapter should contain enough material to get you started.

Why do we need the abstraction provided by measure theory? In your undergraduate probability course, you likely encountered mostly discrete or real-valued random variables. In the former case, we can simply assign to every possible outcome of a random variable a probability; taking expectations is then easy! In the latter case, you probably worked with probability *densities*, i.e.,

$$\text{Prob}(X \in [a,b]) = \int_a^b p_X(x)\,dx, \qquad \mathbb{E}(X) = \int_{-\infty}^{\infty} x\,p_X(x)\,dx, \tag{1.0.1}$$

where p_X is the density of the real-valued random variable X. Though both of these are special cases of the general measure-theoretic framework, one can often easily make do without the general theory.

Unfortunately, this simple form of probability theory will simply not do for our purposes. For example, consider the Wiener process W_t. The map $t \mapsto W_t$ is not a random number, but a random *sample path*. If we wanted to describe the law of this random path by a probability density, the latter would be a function on the space of continuous paths. But how can we then make sense of expressions such as eq. (1.0.1)? What does it mean to integrate a function over the space of continuous paths, or to take limits of such functions? Such questions have to be resolved before we can move on.

1.1 Probability spaces and events

To build a probability model, we need at least three ingredients. We need to know:

- What are all the things that could possibly happen?
- What sensible yes-no questions can we ask about these things?
- For any such question, what is the probability that the answer is yes?

The first point on the agenda is formalized by specifying a set Ω. Every element $\omega \in \Omega$ symbolizes one possible fate of the model.

Example 1.1.1. A coin flip could be modelled by $\Omega = \{\text{heads}, \text{tails}\}$, a roll of a single die by $\Omega = \{1, 2, 3, 4, 5, 6\}$, a roll of two dice (or of the same die twice in a row!) by $\Omega = \{(1, 1), (1, 2), \ldots, (1, 6), (2, 1), \ldots, (6, 6)\}$.

Example 1.1.2. The position of the particle in a fluid could be modelled by $\Omega = \mathbb{R}^3$.

Example 1.1.3. The random motion of the particle in a fluid could be modelled using $\Omega = C([0, \infty[; \mathbb{R}^3)$, the space of $\mathbb{R}^3$-valued continuous functions of time $[0, \infty[$.

Once we have specified Ω, any yes-no question is represented by the subset of Ω consisting of those $\omega \in \Omega$ for which the answer is yes.

Example 1.1.4. Suppose we throw a die, so $\Omega = \{1, 2, 3, 4, 5, 6\}$. The question *did we throw a three?* is represented by the subset $\{3\}$, *did we throw a three or a six?* by $\{3, 6\}$, *did we throw an even number?* by $\{2, 4, 6\}$, etc. You get the picture.

We need to specify what yes-no questions make sense. We will collect all sensible yes-no questions in a set $\mathcal{F}$, i.e., $\mathcal{F}$ is a set of subsets of Ω. Not every such $\mathcal{F}$ qualifies, however. Suppose that $A, B \subset \Omega$ are sensible. Then *A and B?* and *A or B?* should also be sensible questions to ask.[1] Convince yourself that *A and B?* is precisely the question $A \cap B$, and *A or B?* is the question $A \cup B$. Similarly, if $A \subset \Omega$ is a sensible question, its complement *not A?* should also make sense; the latter is clearly equivalent to $A^c \equiv \Omega \backslash A$. Finally, the deep question Ω (*is anything true?*) should always be allowed. Except for a small addition, we have grasped the right concept.

Definition 1.1.5. A *σ-algebra* $\mathcal{F}$ is a collection of subsets of Ω such that

1. If $A_n \in \mathcal{F}$ for countable n, then $\bigcup_n A_n \in \mathcal{F}$.
2. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.
3. $\Omega \in \mathcal{F}$.

An element $A \in \mathcal{F}$ is called an *($\mathcal{F}$-)measurable set* or an *event*.

[1] A curiosity: in quantum mechanics, this is not true—this is a major difference between quantum probability and classical probability. Now that you have read this footnote, be sure to forget it.

The second and third condition are exactly as discussed above. In the first condition, we have allowed not only A *or* B*?*, but also A_1 *or* A_2 *or* A_3 *or* $\ldots$*?*, as long as the number of questions A_n are countable. This is desirable: suppose, for example, that $\Omega = \mathbb{N}$, and that $\{n\} \in \mathcal{F}$ for any $n \in \mathbb{N}$ (*is it three? is it six?* $\ldots$); then it would be a little strange if we could not answer the question $\{2n : n \in \mathbb{N}\}$ (*is it an even number?*). Note that the fact that $\mathcal{F}$ is closed under countable intersections (*and*s) follows from the definition: after all, $\bigcap_n A_n = (\bigcup_n A_n^c)^c$.

Example 1.1.6. Let Ω be any set. Then the *power set* $\mathcal{F} = \{A : A \subset \Omega\}$ (the collection of all subsets of Ω) is a σ-algebra.

We can make more interesting σ-algebras as follows.

Definition 1.1.7. Let $\{A_i\}$ be a (not necessarily countable) collection of subsets of Ω. Then $\mathcal{F} = \sigma\{A_i\}$ denotes the smallest σ-algebra that contains every set A_i, and is called the σ-algebra *generated by* $\{A_i\}$.

It is perhaps not entirely obvious that $\sigma\{A_i\}$ exists or is uniquely defined. But note that the power set contains all $A_i \subset \Omega$, so that there exists at least one σ-algebra that contains all A_i. For uniqueness, note that if $\{\mathcal{F}_j\}$ is a (not necessarily countable) collection of σ-algebras, then $\bigcap_j \mathcal{F}_j$ is also a σ-algebra (check this!) So $\sigma\{A_i\}$ is uniquely defined as the intersection of all σ-algebras that contain all A_i.

Example 1.1.8. Let $\Omega = \{1, 2, 3, 4, 5, 6\}$. Then the σ-algebra generated by $\{1\}$ and $\{4\}$ is $\sigma\{\{1\}, \{4\}\} = \{\varnothing, \{1\}, \{4\}, \{1\}^c, \{4\}^c, \{1, 4\}, \{1, 4\}^c, \Omega\}$. Interpretation: if I can answer the questions *did we throw a one?* and *did we throw a four?*, then I can immediately answer all the questions in $\sigma\{\{1\}, \{4\}\}$. We think of $\sigma\{\{1\}, \{4\}\}$ as encoding the information contained in the observation of $\{1\}$ and $\{4\}$.

This example demonstrates that even if our main σ-algebra is large—in the example of throwing a die, one would normally choose the σ-algebra $\mathcal{F}$ of all sensible questions to be the power set—it is natural to use subalgebras of $\mathcal{F}$ to specify what (limited) information is actually available to us from making certain observations. This idea is very important and will come back again and again.

Example 1.1.9. Let Ω be a topological space. Then $\sigma\{A \subset \Omega : A \text{ is an open set}\}$ is called the *Borel σ-algebra* on Ω, denoted as $\mathcal{B}(\Omega)$.

When we work with continuous spaces, such as $\Omega = \mathbb{R}$ or $\Omega = C([0, \infty[; \mathbb{R}^3)$ (with its natural topology of uniform convergence on compact sets), we will usually choose the σ-algebra $\mathcal{F}$ of sensible events to be the Borel σ-algebra.

Remark 1.1.10. This brings up a point that has probably puzzled you a little. What is all this fuss about "sensible" events (yes-no questions)? If we think of Ω as the set of all possible fates of the system, then why should any event $A \subset \Omega$ fail to be sensible? In particular, why not always choose $\mathcal{F}$ to be the power set? The answer to this question might not be very satisfying. The fact of the matter is that, as was learned the hard way, it is essentially impossible to build a consistent theory if $\mathcal{F}$ contains too many sets. We will come back to this very briefly below and give a

slightly more satisfying answer. This also provides an excuse for another potentially puzzling aspect: why do we only allow countable unions in the definition of the σ-algebra, not uncountable unions? Note that if $\mathcal{F}$ had to be closed under uncountable unions, and contained all individual points of Ω (surely a desirable state of affairs), then $\mathcal{F}$ would be the power set and we would be in trouble. If you are interested in this sort of thing, you will find plenty written about this in the literature. We will accept it as a fact of life, however, that the power set is too large; fortunately, the Borel σ-algebra is an extremely rich object and is more than sufficient for most purposes.

It remains to complete the final point on our agenda: we need to assign a probability to every event in $\mathcal{F}$. Of course, this has to be done in a consistent way. If A and B are two mutually exclusive events ($A \cap B = \varnothing$), then it must be the case that the probability of *A or B?* is the sum of the individual probabilities. This leads to the following definition, which should look very familiar.

Definition 1.1.11. A *probability measure* is a map $\mathbb{P} : \mathcal{F} \to [0,1]$ such that

1. For countable $\{A_n\}$ s.t. $A_n \cap A_m = \varnothing$ for $n \neq m$, $\mathbb{P}(\bigcup_n A_n) = \sum_n \mathbb{P}(A_n)$.
2. $\mathbb{P}(\varnothing) = 0,\ \mathbb{P}(\Omega) = 1$.

The first property is known as *countable additivity*. It is fundamental to the interpretation of the theory but also to its mathematical structure: this property will allow us to take limits, and we will spend a lot of time taking limits in this course.

Definition 1.1.12. A *probability space* is a triple $(\Omega, \mathcal{F}, \mathbb{P})$.

The simplest examples are the point mass and a finite probability space.

Example 1.1.13. Let Ω be any set and $\mathcal{F}$ be any σ-algebra. Fix some $\tilde{\omega} \in \Omega$. Define $\mathbb{P}$ as follows: $\mathbb{P}(A) = 1$ if $\tilde{\omega} \in A$, and $\mathbb{P}(A) = 0$ otherwise. Then $\mathbb{P}$ is a probability measure, called the *point mass* on $\tilde{\omega}$. Intuitively, this corresponds to the situation where the fate $\tilde{\omega}$ always happens ($\mathbb{P}$ is a "deterministic" measure).

Example 1.1.14. Let Ω be a finite set, and $\mathcal{F}$ be the power set of Ω. Then any probability measure on Ω can be constructed as follows. First, specify for every point $\omega \in \Omega$ a probability $\mathbb{P}(\{\omega\}) \in [0,1]$, such that $\sum_{\omega \in \Omega} \mathbb{P}(\{\omega\}) = 1$. We can now extend this map $\mathbb{P}$ to all of $\mathcal{F}$ by using the additivity property: after all, any subset of Ω is the disjoint union of a finite number of sets $\{\omega\}$, so we must have $\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\})$.

This example demonstrates a basic idea: in order to define $\mathbb{P}$, it is not necessary to go through the effort of specifying $\mathbb{P}(A)$ for every element $A \in \mathcal{F}$; it is usually enough to specify the measure on a much smaller class of sets $\mathcal{G} \subset \mathcal{F}$, and if $\mathcal{G}$ is large enough there will exist only one measure that is consistent with the information provided. For example, if Ω is finite, then $\mathcal{G} = \{\{\omega\} : \omega \in \Omega\}$ is a suitable class.

When Ω is continuous, however, specifying the probability of each point $\{\omega\}$ is clearly not enough. Consider, for example, the uniform distribution on $[0,1]$: the probability of any isolated point $\{\omega\}$ should surely be zero! Nonetheless a similar idea holds also in this case, but we have to choose $\mathcal{G}$ a little more carefully. For the

case when $\Omega = \mathbb{R}$ and $\mathcal{F} = \mathcal{B}(\mathbb{R})$, the Borel σ-algebra on $\mathbb{R}$, the appropriate result is stated as the following theorem. The proof of this theorem is far beyond our scope, though it is well worth the effort; see [Bil86, Theorem 12.4].

Theorem 1.1.15. *Let $\mathbb{P}$ be a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Then the function $F(x) = \mathbb{P}(]-\infty, x])$ is nondecreasing, right-continuous, and $F(x) \to 0$ as $x \to -\infty$, $F(x) \to 1$ as $x \to \infty$. Conversely, for any function $F(x)$ with these properties, there exists a unique probability measure $\mathbb{P}$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $F(x) = \mathbb{P}(]-\infty, x])$.*

You must have encountered the function $F(x)$ in your introductory probability course: this is the cumulative distribution function (CDF) for the measure $\mathbb{P}$ on $\mathbb{R}$. Theorem 1.1.15 forms a link between introductory probability, which centers around objects such as $F(x)$, and more advanced probability based on measure spaces.

In this course we will never need to construct probability measures directly on more complicated spaces than $\mathbb{R}$. As we will see, various techniques allow us to construct more complicated probability spaces from simpler ones.

Example 1.1.16. The simple Gaussian probability space with mean $\mu \in \mathbb{R}$ and variance $\sigma > 0$ is given by $(\Omega, \mathcal{F}, \mathbb{P})$ with $\Omega = \mathbb{R}$, $\mathcal{F} = \mathcal{B}(\mathbb{R})$, and $\mathbb{P}$ is constructed through Theorem 1.1.15 using $F(x) = \frac{1}{2} + \frac{1}{2}\,\mathrm{erf}((x-\mu)/\sigma\sqrt{2})$.

Remark 1.1.17. We can now say a little more about the discussion in remark 1.1.10. Suppose we took $\Omega = \mathbb{R}$, say, and we took $\mathcal{F}$ to be the power set. What would go wrong? It turns out that *there do not exist any probability measures on the power set of $\mathbb{R}$ such that $\mathbb{P}(\{x\}) = 0$ for all $x \in \mathbb{R}$.* This is shown by Banach and Kuratowski [BK29]; for more information, see [Dud02, Appendix C] or [Bir67, sec. XI.7]. This means that if we wanted to work with the power set, the probability mass could at best concentrate only on a countable number of points; but then we might as well choose Ω to be the set of those points, and discard the rest of $\mathbb{R}$. The proof of Banach and Kuratowski assumes the continuum hypothesis, so might be open to some mathematical bickering; but at the end of the day it seems pretty clear that we are not going to be able to do anything useful with the power set. For us, the case is now closed.

1.2 Some elementary properties

Now that our basic definitions are in place, we can start pushing them around. The following results are extremely simple and extremely useful; they are direct consequences of our definitions and some set manipulation gymnastics! If you have never seen these before, take a piece of scratch paper and try to prove them yourself.

Lemma 1.2.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.*

1. $A \in \mathcal{F} \Longrightarrow \mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
2. $A, B \in \mathcal{F}$, $A \subset B \Longrightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$.
3. $\{A_n\} \subset \mathcal{F}$ *countable* $\Longrightarrow \mathbb{P}(\bigcup_n A_n) \leq \sum_n \mathbb{P}(A_n)$.
4. $A_1 \subset A_2 \subset \cdots \in \mathcal{F} \Longrightarrow \lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(\bigcup_n A_n)$.

5. $A_1 \supset A_2 \supset \cdots \in \mathcal{F} \Longrightarrow \lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(\bigcap_n A_n)$.

Proof.

1. $\Omega = A \cup A^c$ and $A \cap A^c = \varnothing$, so $1 = \mathbb{P}(\Omega) = \mathbb{P}(A) + \mathbb{P}(A^c)$.
2. $B = A \cup (B\backslash A)$ and $A \cap (B\backslash A) = \varnothing$, so $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B\backslash A) \geq \mathbb{P}(A)$.
3. Assume without loss of generality that $n \in \mathbb{N}$ ($\{A_n\}$ is a sequence). We construct sets $\{B_n\}$ that are disjoint, i.e., $B_n \cap B_m = \varnothing$ for $m \neq n$, but such that $\bigcup_k A_k = \bigcup_k B_k$: choose $B_1 = A_1$, $B_2 = A_2\backslash A_1$, $B_3 = A_3\backslash(A_1 \cup A_2)$, ... But note that $B_k \subset A_k$ for any k. Hence $\mathbb{P}(B_k) \leq \mathbb{P}(A_k)$, and we obtain
$$\mathbb{P}\left(\bigcup_k A_k\right) = \mathbb{P}\left(\bigcup_k B_k\right) = \sum_k \mathbb{P}(B_k) \leq \sum_k \mathbb{P}(A_k).$$
4. Write $B_1 = A_1$, $B_k = A_k\backslash A_{k-1}$, so $\mathbb{P}(B_k) = \mathbb{P}(A_k) - \mathbb{P}(A_{k-1})$. The B_k are disjoint and their union is the union of the A_k, so we obtain
$$\mathbb{P}\left(\bigcup_k A_k\right) = \mathbb{P}\left(\bigcup_k B_k\right) = \mathbb{P}(A_1) + \sum_{k=2}^{\infty}\{\mathbb{P}(A_k) - \mathbb{P}(A_{k-1})\} = \lim_{k\to\infty} \mathbb{P}(A_k).$$
5. Use $\bigcap_n A_n = (\bigcup_n A_n^c)^c$ and the previous result. □

The following simple corollary is worth emphasizing.

Corollary 1.2.2. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probab. space, and let $\{A_n\} \subset \mathcal{F}$ be countable.*

1. *If $\mathbb{P}(A_n) = 0$ for all n, then $\mathbb{P}(\bigcup_n A_n) = 0$.*
2. *If $\mathbb{P}(A_n) = 1$ for all n, then $\mathbb{P}(\bigcap_n A_n) = 1$.*

In words: if every event A_n has zero probability of happening, then with unit probability none of these events happen. If every event A_n happens with unit probability, then the probability that all these events occur simultaneously is one.

Remark 1.2.3. This may seem a tautology, but it is nice to see that our intuition is faithfully encoded in the mathematics. More importantly, however, note that the statement is not true for uncountable families $\{A_n\}$. For example, under the uniform distribution on $[0, 1]$, any individual outcome $\{x\}$ has zero probability of occuring. However, the probability that one of these outcomes occurs is $\mathbb{P}([0, 1]) = 1$!

Let us now introduce another useful concept. Suppose that $\{A_n\} \subset \mathcal{F}$ is a sequence of measurable sets. We would like to know: what is the set of points $\omega \in \Omega$ which are an element of *infinitely many* of the A_n? Sometimes this set is denoted as $\{\omega \in \Omega : \omega \in A_n \text{ i.o.}\}$, where i.o. stands for infinitely often. We will find that this concept is very useful in proving convergence of a sequence of random variables.

Let us characterize this set. For some $\omega \in \Omega$, clearly $\omega \in A_n$ infinitely often if and only if for any n, there is an $N(n, \omega) \geq n$ such that $\omega \in A_{N(n,\omega)}$. That is,
$$\omega \in A_n \text{ i.o.} \quad \Longleftrightarrow \quad \omega \in \bigcup_{k\geq n} A_k \ \forall n \quad \Longleftrightarrow \quad \omega \in \bigcap_{n\geq 1} \bigcup_{k\geq n} A_k.$$

The rightmost set is called $\limsup A_k$, in analogy with the limit superior of a sequence of numbers. We have thus established:

$$\{\omega \in \Omega : \omega \in A_n \text{ i.o.}\} = \bigcap_{n\geq 1} \bigcup_{k\geq n} A_k = \limsup A_k.$$

This also proves measurability, i.e., $\{\omega \in \Omega : \omega \in A_n \text{ i.o.}\} \in \mathcal{F}$ (why?). This was not entirely obvious to begin with!

We can now prove various results for this set; for example, you should prove that $\mathbb{P}(\limsup A_k) \geq \limsup \mathbb{P}(A_k)$. The most useful result, however, is the following.

Lemma 1.2.4 (Borel-Cantelli). *If* $\sum_n \mathbb{P}(A_n) < \infty$, *then* $\mathbb{P}(\limsup A_n) = 0$.

Proof. Simply use lemma 1.2.1:

$$\mathbb{P}(\limsup A_n) = \mathbb{P}\left(\bigcap_{n\geq 1} \bigcup_{k\geq n} A_k\right) = \lim_{n\to\infty} \mathbb{P}\left(\bigcup_{k\geq n} A_k\right) \leq \lim_{n\to\infty} \sum_{k\geq n} \mathbb{P}(A_k) = 0,$$

where we have used that $\bigcup_{k\geq n} A_k$ is a nonincreasing sequence of sets. □

1.3 Random variables and expectation values

The next most important ingredient in probability theory is the random variable. If $(\Omega, \mathcal{F}, \mathbb{P})$ describes all possible fates of the system as a whole and their probabilities, then random variables describe concrete observations that we can make on the system. That is, suppose that we have a measurement apparatus that returns an element in some set S; for example, it could measure a real number (such as a measurement of distance using a ruler), a point on the circle (measuring an angle), a point in a finite set, and entire trajectory, ... The outcome of such a measurement is described by specifying what value it takes for every possible fate of the system $\omega \in \Omega$.

Definition 1.3.1. An *($\mathcal{F}$-)measurable function* is a map $f : \Omega \to S$ from $(\Omega, \mathcal{F})$ to $(S, \mathcal{S})$ such that $f^{-1}(A) \equiv \{\omega \in \Omega : f(\omega) \in A\} \in \mathcal{F}$ for every $A \in \mathcal{S}$. If $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $(S, \mathcal{S})$ is a measurable space, then a measurable function $f : \Omega \to S$ is called an *S-valued random variable*. A real-valued random variable ($S = \mathbb{R}$, $\mathcal{S} = \mathcal{B}(\mathbb{R})$) is often just called a *random variable*.

The notion of measurability is fundamental to our interpretation of the theory. Suppose we have a measurement apparatus that returns a real number; this is described by a random variable $X : \Omega \to \mathbb{R}$. At the very least, our model should be able to answer the question: if we perform such a measurement, what is the probability of observing a measurement outcome in some set $A \in \mathcal{B}(\mathbb{R})$? Clearly this probability is precisely $\mathbb{P}(X^{-1}(A))$; for this expression to make sense, X has to be measurable.

Remark 1.3.2 (Common notational conventions). We will often overload notation in obvious but aesthetically pleasing ways. For example, the probability that the random variable $X : \Omega \to S$ takes a value in $A \in \mathcal{S}$ could be denoted by $\mathbb{P}(X \in A)$; technically, of course, we should write $\mathbb{P}(\{\omega \in \Omega : X(\omega) \in A\})$. Similarly we will

encounter notation such as $\mathbb{P}(X > 0)$, $\mathbb{P}(|X_n| > \varepsilon \text{ i.o.})$, etc. Such notation is very intuitive, but keep in mind that this is actually short-hand notation for well-defined mathematical objects: the probabilities of certain events in $\mathcal{F}$.

We will take another notational liberty. If we make some statement, for example, if we claim that $X \in A$ (i.e., we claim to have proved that $X \in A$) or that $|X_n| > \varepsilon$ infinitely often, we generally mean that that statement is true with probability one, e.g., $\mathbb{P}(X \in A) = 1$ or $\mathbb{P}(|X_n| > \varepsilon \text{ i.o.}) = 1$. If we wanted to be precise, we would say explicitly that the statement holds *almost surely* (abbreviated as *a.s.*). Though sets of probability zero do not always play a negligible role (see section 2.4), we are ultimately only interested in proving results with unit probability, so it is convenient to interpret all intermediate statements as holding with probability one.

Now you might worry (as well you should!) that this sort of sloppiness could get us in big trouble; but we claim that as long as we make only *countably* many almost sure statements, we have nothing to worry about. You should revisit Corollary 1.2.2 at this point and convince yourself that this logic is air-tight.

It is not always entirely trivial to prove that a map is measurable. The following simple facts are helpful and not difficult to prove; see, for example, [Wil91, Ch. 3].

Lemma 1.3.3. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(S, \mathcal{S})$ be a measurable space.*

1. *If $h : \Omega \to S$ and $f : S \to S'$ are measurable, then $f \circ h$ is measurable.*
2. *If $\{h_n\}$ is a sequence of measurable functions $h_n : \Omega \to S$, then $\inf_n h_n$, $\sup_n h_n$, $\liminf_n h_n$, and $\limsup_n h_n$ are measurable.*
3. *If $h_1, h_2 : \Omega \to \mathbb{R}$ are measurable, then so are $h_1 + h_2$ and $h_1 h_2$.*
4. *If Ω is a topological space with its Borel σ-algebra $\mathcal{F} = \mathcal{B}(\Omega)$, then any continuous function $h : \Omega \to \mathbb{R}$ is measurable.*

The following idea is very important:

Definition 1.3.4. Let Ω be a set and $(S, \mathcal{S})$ be a measurable space. Let $\{h_i\}_{i \in I}$ be a (not necessarily countable) collection of maps $h_i : \Omega \to S$. Then $\sigma\{h_i\}$ denotes the smallest σ-algebra on Ω with respect to which every h_i is measurable, and is called the *σ-algebra generated by* $\{h_i\}$. Note that $\sigma\{h_i\} = \sigma\{h_i^{-1}(A) : A \in \mathcal{S},\ i \in I\}$.

One could use this as a method to generate a σ-algebra on Ω, if we did not have one to begin with, starting from the given σ-algebra $\mathcal{S}$. However, usually this concept is used in a different way. We start with a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and consider some collection $\{X_i\}$ of random variables (which are already $\mathcal{F}$-measurable). Then $\sigma\{X_i\} \subset \mathcal{F}$ is the sub-σ-algebra of $\mathcal{F}$ which contains precisely those yes-no questions that can be answered by measuring the X_i. In this sense, $\sigma\{X_i\}$ represents the information that is obtained by measuring the random variables X_i.

Example 1.3.5. Suppose we toss two coins, so we model $\Omega = \{HH, HT, TH, TT\}$, $\mathcal{F}$ is the power set of Ω, and we have some measure $\mathbb{P}$ which is irrelevant for this discussion. Suppose we only get to observe the outcome of the first coin flip, i.e., we

see the random variable $X(HH) = X(HT) = 1$, $X(TH) = X(TT) = 0$. Then $\sigma\{X\} = \{\varnothing, \{HH, HT\}, \{TH, TT\}, \Omega\} \subset \mathcal{F}$ contains only those yes-no questions which can be answered from our knowledge of the outcome of the first coin flip (e.g., the event $\{HH, HT\}$ is *did the first coin come up heads?*).

The converse of this idea is equally important. Suppose that we are working on $(\Omega, \mathcal{F}, \mathbb{P})$, and that $\mathcal{G} \subset \mathcal{F}$ is some sub-σ-algebra representing a limited amount of information. We ask: when is the knowledge of the information in $\mathcal{G}$ sufficient to determine the outcome of some random variable X? It is easy to see that the answer to every question we can ask about X can be determined from the available information $\mathcal{G}$ if and only if X is $\mathcal{G}$-measurable (why?) The following lemma, for a special case, suggests how we could think intuitively about this idea.

Lemma 1.3.6. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X_1, \dots, X_n$ and X be real-valued random variables. Suppose that X is $\sigma\{X_1, \dots, X_n\}$-measurable. Then there exists a measurable map $f : \mathbb{R}^n \to \mathbb{R}$ such that $X = f(X_1, \dots, X_n)$.*

Partial proof. Let us prove the result for the case where Ω is a finite set and $\mathcal{F}$ is the power set. The general proof proceeds along the same lines, see [Bil86, Theorem 20.1].

When Ω is a finite set, X and $X_1, \dots, X_n$ can each only take a finite number of values; let us write $\Xi, \Xi_i \subset \mathbb{R}$ for the possible values of X and X_i, respectively. We can now consider $X_* = (X_1, \dots, X_n)$ as an $\Xi_1 \times \cdots \times \Xi_n$-valued random variable, and we would like to prove that $X = f(X_*)$ for some function $f : \Xi_1 \times \cdots \times \Xi_n \to \Xi$.

As X is $\sigma\{X_1, \dots, X_n\}$-measurable, we have $X^{-1}(x) \in \sigma\{X_1, \dots, X_n\}$ for any $x \in \Xi$. It is not difficult to convince yourself that $\sigma\{X_1, \dots, X_n\}$ consists of the empty set $\varnothing$ and of all sets $X_*^{-1}(A)$ with $A \subset \Xi_1 \times \cdots \times \Xi_n$. Hence for every x, there is an A_x such that $X^{-1}(x) = X_*^{-1}(A_x)$, and we have $A_x \cap A_y = \varnothing$ for $x \neq y$ and $\bigcup_x A_x = \Xi_1 \times \cdots \times \Xi_n$. We can now define the function f uniquely by setting $f(\xi) = x$ for all $\xi \in A_x$. □

We will not need this lemma in the rest of this course; it is included here to help you form an intuition about measurability and generated σ-algebras. The point is that if $\{X_i\}$ is a collection of random variables and X is $\sigma\{X_i\}$-measurable, then you should think of X as being a function of the X_i. It is possible to prove analogs of lemma 1.3.6 for most situations of interest (even when the collection $\{X_i\}$ is not finite), if one so desires, but there is rarely a need to do so.

The rest of this section is devoted to the concept of *expectation*. For a random variable that takes a finite number of values, you know very well what this means: it is the sum of the values of the random variable weighted by their probabilities.

Definition 1.3.7. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A *simple random variable* $X : \Omega \to \mathbb{R}$ is a random variable that takes only a finite number of values, i.e., $X(\Omega) = \{x_1, \dots, x_n\}$. Its expectation is defined as $\mathbb{E}(X) = \sum_{k=1}^n x_k \, \mathbb{P}(X = x_k)$.

Remark 1.3.8. Sometimes we will be interested in multiple probability measures on the same σ-algebra $\mathcal{F}$ ($\mathbb{P}$ and $\mathbb{Q}$, say). The notation $\mathbb{E}(X)$ can then be confusing: do we mean $\sum_k x_k \, \mathbb{P}(X = x_k)$ or $\sum_k x_k \, \mathbb{Q}(X = x_k)$? Whenever necessary, we will denote the former by $\mathbb{E}_{\mathbb{P}}$ and the latter by $\mathbb{E}_{\mathbb{Q}}$ to avoid confusion. Usually, however, we will be working on some fixed probability space and there will be only one measure of interest $\mathbb{P}$; in that case, it is customary to write $\mathbb{E}(X)$ to lighten the notation.

We want to extend this definition to general random variables. The simplest extension is to the case where X does not take a finite number of values, but rather a countable number of values. This appears completely trivial, but there is an issue here of the elementary calculus type: suppose that X takes the values $\{x_k\}_{k\in\mathbb{N}}$ and we define $\mathbb{E}(X) = \sum_{k=1}^{\infty} x_k \, \mathbb{P}(X = x_k)$. It is not obvious that this sum is well behaved: if x_k is an alternating sequence, it could well be that the series $\sum_{k=1}^{n} x_k \, \mathbb{P}(X = x_k)$ is not absolutely convergent and the expectation would thus depend on the order of summation! Clearly that sort of thing should not be allowed. To circumvent this problem we introduce the following definition, which holds generally.

Definition 1.3.9. Let us define $X^+ = \max(X, 0)$ and $X^- = -\min(X, 0)$, so that $X = X^+ - X^-$. The expectation $\mathbb{E}(X)$ is defined only if either $\mathbb{E}(X^+) < \infty$ or $\mathbb{E}(X^-) < \infty$. If this is the case, then by definition $\mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X^-)$.

As such, we should concentrate on defining $\mathbb{E}(X)$ for nonnegative X. We have got this down for simple random variables and for random variables with countable values; what about the general case? The idea here is very simple. For any nonnegative random variable X, we can find a sequence X_n of *simple* random variables that converges to X; actually, it is most convenient to choose X_n to be a nondecreasing sequence $X_n \nearrow X$ so that $\mathbb{E}(X_n)$ is guaranteed to have a limit (why?).

Definition 1.3.10. Let X by any nonnegative random variable. Then we define the expectation $\mathbb{E}(X) = \lim_{n\to\infty} \mathbb{E}(X_n)$, where X_n is any nondecreasing sequence of simple random variables that converges to X.

It remains to prove (a) that we can find such a sequence X_n; and (b) that any such sequence gives rise to the same value for $\mathbb{E}(X)$. Once these little details are established, we will be convinced that the definition of $\mathbb{E}(X)$ makes sense. If you are already convinced, read the following remark and then skip to the next section.

Remark 1.3.11. The idea of approximating a function by a piecewise constant function, then taking limits should look very familiar—remember the Riemann integral? In fact, the expectation which we have constructed really is a type of integral, the *Lebesgue integral* with respect to the measure $\mathbb{P}$. It can be denoted in various ways:

$$\mathbb{E}(X) \equiv \int X(\omega) \, \mathbb{P}(d\omega) \equiv \int X \, d\mathbb{P}.$$

Unlike the Riemann integral we can use the Lebesgue integral to integrate functions on very strange spaces: for example, as mentioned at the beginning of the chapter, we can integrate functions on the space of continuous paths—provided that we can construct a suitable measure $\mathbb{P}$ on this space.

When $\Omega = \mathbb{R}^d$, $\mathcal{F} = \mathcal{B}(\mathbb{R}^d)$ and with a suitable choice of measure μ (instead of $\mathbb{P}$), the Lebesgue integral can actually serve as a generalization of the Riemann integral (it is a generalization because the Riemann integral can only integrate continuous functions, whereas the Lebesgue integral can integrate measurable functions). The *Lebesgue measure* μ, however, is not a probability measure: it satisfies all the conditions of Definition 1.1.11 except $\mu(\Omega) = 1$ (as $\mu(\mathbb{R}^d) = \infty$). This does not change much, except that we can obviously not interpret μ probabilistically.

Let us tie up the loose ends in our definition of the expectation.

Lemma 1.3.12. *Let X be a nonnegative random variable. Then there exists a nondecreasing sequence of simple random variables X_n such that $X_n \nearrow X$.*

Proof. Define X_n as

$$X_n(\omega) = \begin{cases} 0 & \text{if } X(\omega) = 0, \\ (k-1)2^{-n} & \text{if } (k-1)2^{-n} < X(\omega) \le k2^{-n}, \;\; k = 1, \dots, n2^n, \\ n & \text{if } X(\omega) > n. \end{cases}$$

(Why is X_n measurable?) Clearly $X_n \nearrow X$, and we are done. □

Lemma 1.3.13. *Let $X \ge 0$, and let $\{X_n\}$ and $\{\tilde{X}_n\}$ be two sequences of simple random variables s.t. $X_n \nearrow X$ and $\tilde{X}_n \nearrow X$. Then $\lim_{n\to\infty} \mathbb{E}(X_n) = \lim_{n\to\infty} \mathbb{E}(\tilde{X}_n)$.*

Proof. It suffices to prove that $\mathbb{E}(\tilde{X}_k) \le \lim_{n\to\infty} \mathbb{E}(X_n)$ for any k. After all, this implies that $\lim_{k\to\infty} \mathbb{E}(\tilde{X}_k) \le \lim_{n\to\infty} \mathbb{E}(X_n)$, and inequality in the reverse direction follows by reversing the roles of X_n and $\tilde{X}_n$. To proceed, note that as $\tilde{X}_k$ is simple, it takes a finite number of values $x_1, \dots, x_\ell$ on the sets $A_i = \tilde{X}_k^{-1}(x_i)$. Define

$$B_i^n = \{\omega \in A_i : X_n(\omega) \ge x_i - \varepsilon\}.$$

Note that $X_n(\omega) \nearrow X(\omega)$ and $\tilde{X}_k(\omega) \le X(\omega)$, so $B_i^n \subset B_i^{n+1} \subset \cdots$ and $\bigcup_n B_i^n = A_i$. By lemma 1.2.1, we have $\mathbb{P}(B_i^n) \nearrow \mathbb{P}(A_i)$. But it is not difficult to see that

$$\mathbb{E}(X_n) \ge \sum_{i=1}^{\ell} (x_i - \varepsilon)\, \mathbb{P}(B_i^n) \quad \Longrightarrow \quad \lim_{n\to\infty} \mathbb{E}(X_n) \ge \mathbb{E}(\tilde{X}_k) - \varepsilon.$$

As this holds for any $\varepsilon > 0$, the statement follows. □

1.4 Properties of the expectation and inequalities

Having defined the expectation, let us first investigate some of its simplest properties. Most of these are trivial for simple random variables and will be well known to you; but can you prove them in the general case?

Lemma 1.4.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, X, Y be random variables whose expectations are assumed to be defined, and $\alpha, \beta \in \mathbb{R}$ are constants.*

1. *If $X = Y$ a.s., then $\mathbb{E}(X) = \mathbb{E}(Y)$.*
2. *If $X \le Y$ a.s., then $\mathbb{E}(X) \le \mathbb{E}(Y)$.*
3. *$\mathbb{E}(\alpha X + \beta Y) = \alpha\mathbb{E}(X) + \beta\mathbb{E}(Y)$ provided the right hand side is not $\infty - \infty$.*
4. *$|\mathbb{E}(X)| \le \mathbb{E}(|X|)$.*
5. *If $\mathbb{E}(X)$ is finite, then X is finite a.s.*
6. *If $X \ge 0$ a.s. and $\mathbb{E}(X) = 0$, then $X = 0$ a.s.*

Proof. The main idea is to prove that these results are true for simple random variables (this is easily verified), then take appropriate limits.

1. First assume $X, Y \geq 0$. Apply lemma 1.3.12 to X, Y; this gives two sequences X_n, Y_n of simple functions with $X_n \nearrow X$, $Y_n \nearrow Y$, and $X_n = Y_n$ a.s. for all n (why?). It is immediate that $\mathbb{E}(X_n) = \mathbb{E}(Y_n)$ for all n, so the result follows by letting $n \to \infty$. Now drop the assumption $X, Y \geq 0$ by considering separately X^+, Y^+ and X^-, Y^-.
2. Same idea.
3. Same idea.
4. Use $-|f| \leq f \leq |f|$ and that $X \leq Y$ implies $\mathbb{E}(X) \leq \mathbb{E}(Y)$.
5. Suppose X is not finite a.s.; then on some set $A \in \mathcal{F}$ with $\mathbb{P}(A) > 0$ we have $X = \infty$ or $-\infty$ (we can not have both, as then $\mathbb{E}(X)$ would not be defined). It follows from the definition of the expectation that $\mathbb{E}(X^+) = \infty$ or $\mathbb{E}(X^-) = \infty$, respectively (why?).
6. Suppose that $\mathbb{P}(X > 0) > 0$. We claim that there is a $\varepsilon > 0$ s.t. $\mathbb{P}(X > \varepsilon) > 0$. Indeed, the sets $A_\varepsilon = \{\omega \in \Omega : X(\omega) > \varepsilon\}$ increase in size with decreasing ε, so $\mathbb{P}(A_\varepsilon) \nearrow \mathbb{P}(A_0) = \mathbb{P}(X > 0) > 0$ (remember lemma 1.2.1?), and thus there must exist a positive ε with $\mathbb{P}(A_\varepsilon) > 0$. But then $\mathbb{E}(X) \geq \mathbb{E}(\varepsilon I_{A_\varepsilon}) = \varepsilon \mathbb{P}(A_\varepsilon) > 0$ (here $I_A(\omega) = 1$ if $\omega \in A$, 0 otherwise) which contradicts the assumption. □

Next, let us treat two elementary inequalities: Chebyshev's inequality (often called Markov's inequality) and Jensen's inequality. These inequalities are *extremely useful*: do not leave home without them! In the following, we will often use the notation

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{otherwise,} \end{cases} \qquad A \in \mathcal{F}.$$

The function I_A is called the *indicator* or *characteristic function* on A.

Proposition 1.4.2. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X be a random variable.*

1. ***(Chebyshev/Markov)** For any $\alpha > 0$, we have*

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

2. ***(Jensen)** Let $g(x)$ be a real-valued convex function (such a function is always measurable) and let $\mathbb{E}(X)$ be finite. Then $\mathbb{E}(g(X)) \geq g(\mathbb{E}(X))$.*

Proof. To prove Chebyshev's inequality, let us define $A = \{\omega \in \Omega : |X(\omega)| \geq \alpha\}$. Clearly $|X(\omega)| \geq \alpha I_A(\omega)$, so $\mathbb{E}(|X|) \geq \alpha \mathbb{E}(I_A) = \alpha \mathbb{P}(A)$.

For Jensen's inequality, note that $g(x)$ is continuous (by convexity), so it is measurable. As g is convex, there is a line $f(x) = ax + b$ such that $f(\mathbb{E}(X)) = g(\mathbb{E}(X))$ and $f \leq g$ everywhere. Thus $g(\mathbb{E}(X)) = f(\mathbb{E}(X)) = \mathbb{E}(f(X)) \leq \mathbb{E}(g(X))$. □

Chebyshev's inequality allows us to bound the *tail* of a random variable: it says that if the expectation of a nonnegative random variable X is finite, then X will rarely take very large values. Though the bound is quite crude (in specific situations much tighter bounds on the tails are possible), it is often very effective. Jensen's inequality

is quite fundamental; it says, for example, something that you know very well: the variance $\mathrm{var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$ is always nonnegative (as x^2 is convex).

Recall that the expectation of X is defined when $\mathbb{E}(X^+) < \infty$ or $\mathbb{E}(X^-) < \infty$. The most useful case, however, is when both these quantities are finite: i.e., when $\mathbb{E}(|X|) = \mathbb{E}(X^+) + \mathbb{E}(X^-) < \infty$. In this case, X is said to be *integrable*. It is not necessarily true, e.g., that an integrable random variable has a finite variance; we need to require a little more for this. In fact, there is a hierarchy of regularity conditions.

Definition 1.4.3. For a random variable X and $p \geq 1$, let $\|X\|_p = (\mathbb{E}(|X|^p))^{1/p}$. A random variable with $\|X\|_1 < \infty$ is called *integrable*, with $\|X\|_2 < \infty$ *square integrable*. A random variable is called *bounded* if there exists $K \in \mathbb{R}$ such that $|X| \leq K$ a.s.; the quantity $\|X\|_\infty$ is by definition the smallest such K.

Remark 1.4.4. Almost all the material which we have discussed until this point has had direct intuitive content, and it is important to understand the intuition and ideas behind these concepts. Integrability conditions are a little less easy to visualize; though they usually have significant implications in the theory (many theorems only hold when the random variables involved satisfy $\|X\|_p < \infty$ for some sufficiently large p), they certainly belong more to the technical side of things. Such matters are unavoidable and, if you are a fan of analysis, can be interesting to deal with in themselves (or a pain in the butt, if you will). As we progress through this course, try to make a distinction for yourself between the conceptual challenges and the technical challenges that we will face (though sometimes these will turn out to be intertwined!)

The spaces $\mathcal{L}^p = \mathcal{L}^p(\Omega, \mathcal{F}, \mathbb{P}) = \{X : \Omega \to \mathbb{R};\ \|X\|_p < \infty\}$ play a fundamental role in functional analysis; on $\mathcal{L}^p$, $\|\cdot\|_p$ is almost a norm (it is not a norm, because $\|X\|_p = 0$ implies that $X = 0$ a.s., not $X(\omega) = 0$ for all ω) and $\mathcal{L}^p$ is almost a Banach space; similarly, $\mathcal{L}^2$ is almost a Hilbert space. Functional analytic arguments would give an extra dimension to several of the topics in this course, but as they are not prerequisite for the course we will leave these ideas for you to learn on your own (if you do not already know them). Excellent references are [RS80] and [LL01]. Here we will content ourselves by stating the most elementary results.

Proposition 1.4.5. *Define $\mathcal{L}^p$ as the space of random variables X with $\|X\|_p < \infty$.*

1. *$\mathcal{L}^p$ is linear: if $\alpha \in \mathbb{R}$ and $X, Y \in \mathcal{L}^p$, then $X + Y \in \mathcal{L}^p$ and $\alpha X \in \mathcal{L}^p$.*
2. *If $X \in \mathcal{L}^p$ and $1 \leq q \leq p$, then $\|X\|_q \leq \|X\|_p$ (so $\mathcal{L}^p \subset \mathcal{L}^q$).*
3. ***(Hölder's inequality)** Let $p^{-1} + q^{-1} = 1$. If $X \in \mathcal{L}^p$ and $Y \in \mathcal{L}^q$, then $|\mathbb{E}(XY)| \leq \|X\|_p \|Y\|_q$ (so $XY \in \mathcal{L}^1$).*
4. ***(Minkowski's inequality)** If $X, Y \in \mathcal{L}^p$, then $\|X + Y\|_p \leq \|X\|_p + \|Y\|_p$.*

Proof.

1. Linearity follows immediately from $|x + y|^p \leq (2\,|x| \vee |y|)^p \leq 2^p(|a|^p + |b|^p)$.
2. We would like to prove $\mathbb{E}(|X|^p) = \|X\|_p^p \geq \|X\|_q^p = \mathbb{E}(|X|^q)^{p/q}$. But this follows directly from convexity of $x^{p/q}$ on $[0, \infty[$ and Jensen's inequality.

3. We can restrict ourselves to the case where X and Y are nonnegative and $\|X\|_p > 0$ (why?) For any $A \in \mathcal{F}$, define $\mathbb{Q}(A) = \mathbb{E}(I_A X^p)/\mathbb{E}(X^p)$. Then $\mathbb{Q}$ is also a probability measure on $\mathcal{F}$ (this idea is fundamental, and we will come back to it later on). Define $Z(\omega) = Y(\omega)/X(\omega)^{p-1}$ wherever $X(\omega) > 0$, and $Z(\omega) = 0$ otherwise. But then

$$\mathbb{E}(XY) = \mathbb{E}(X^p)\,\mathbb{E}_{\mathbb{Q}}(Z) \leq \mathbb{E}(X^p)\,(\mathbb{E}_{\mathbb{Q}}(Z^q))^{1/q} = (\mathbb{E}(X^p))^{1-1/q}\,(\mathbb{E}(I_{\{\omega:X(\omega)>0\}}Y^q X/X^{(q-1)(p-1)}))^{1/q} = \|X\|_p\,(\mathbb{E}(Y^q I_{\{\omega:X(\omega)>0\}}))^{1/q} \leq \|X\|_p\|Y\|_q,$$

where we have used that $p^{-1} + q^{-1} = 1$ implies $(q-1)(p-1) = 1$.

4. Let $q^{-1} = 1 - p^{-1}$. We claim that $|X+Y|^{p-1} \in \mathcal{L}^q$. To see this, note that we can write $|X+Y|^{q(p-1)} = |X+Y|^p$ which is integrable by the linearity of $\mathcal{L}^p$. Hence

$$\mathbb{E}(|X+Y|^p) \leq \mathbb{E}(|X|\,|X+Y|^{p-1}) + \mathbb{E}(|Y|\,|X+Y|^{p-1}) \leq (\|X\|_p + \|Y\|_p)\,\|\,|X+Y|^{p-1}\,\|_q = (\|X\|_p + \|Y\|_p)\,\mathbb{E}(|X+Y|^p)^{1/q},$$

using Hölder's inequality, from which the result follows. □

1.5 Limits of random variables

Suppose we have a sequence of random variables X_n. We often want to study limits of such random variables: the sequence X_n converges to some random variable X. We already encountered one such limit in the definition of the expectation, where we meant $X_n \to X$ in the sense that $X_n(\omega) \to X(\omega)$ for every $\omega \in \Omega$. This is not the only way to take limits, however; in fact, there is quite a number of limit concepts for random variables, each of which with its own special properties. At first this may seem like a pure technicality, but the conceptual differences between these limits are very important. We will see, for example, that the selection of the appropriate type of limit is crucial if we wish to define a meaningful stochastic integral (chapter 4). This section introduces the most important concepts which we will need further on.

Definition 1.5.1. Let X be a random variable and $\{X_n\}$ be a sequence of random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

1. $X_n \to X$ *a.s.* if $\mathbb{P}(\{\omega \in \Omega : X_n(\omega) \to X(\omega)\}) = 1$ (why is this event in $\mathcal{F}$?).
2. $X_n \to X$ *in probability* if $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$ as $n \to \infty$ for every $\varepsilon > 0$.
3. $X_n \to X$ *in* $\mathcal{L}^p$ if $\|X_n - X\|_p \to 0$ as $n \to \infty$.
4. $X_n \to X$ *in law* (or *in distribution*, or *weakly*) if $\mathbb{E}(f(X_n)) \to \mathbb{E}(f(X))$ for every bounded continuous function f.

Take a moment to think about these definitions. All of them seem like reasonable ways to characterize the limit of a sequence of random variables—right? Nonetheless all these limit concepts are inequivalent! To give you some feeling for how these concepts work, we will do two things. First, and most importantly, we will prove which type of limit implies which other. Next, we will give illustrative examples of sequences $\{X_n\}$ that converge in one sense, but not in another.

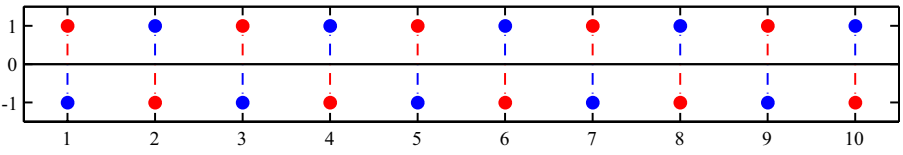

Figure 1.1. The sequence $\{X_n\}$ of example 1.5.3. Shown is the path $n \mapsto X_n(\omega)$ with $\omega = a_1$ (blue) and $\omega = a_2$ (red) for $n = 1, \ldots, 10$. Both paths occur with equal probability.

Proposition 1.5.2. *Let X be a random variable and $\{X_n\}$ be a sequence of random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The following implications hold:*

1. $X_n \to X$ *a.s.* $\implies$ $X_n \to X$ *in probability.*
2. $X_n \to X$ *in* $\mathcal{L}^p$ $\implies$ $X_n \to X$ *in probability.*
3. $X_n \to X$ *in* $\mathcal{L}^p$ $\implies$ $X_n \to X$ *in* $\mathcal{L}^q$ $(q \leq p)$.
4. $X_n \to X$ *in probability* $\implies$ $X_n \to X$ *in law.*

Proof.

1. If $X_n \to X$ a.s. then with unit probability, there is for any $\varepsilon > 0$ an $N(\omega, \varepsilon)$ such that $|X_n(\omega) - X(\omega)| \leq \varepsilon$ for all $n \geq N(\omega, \varepsilon)$. Hence with unit probability, $|X_n - X| > \varepsilon$ only happens for finitely many n, so $\mathbb{P}(|X_n - X| > \varepsilon \text{ i.o.}) = 0$. It remains to note that $\mathbb{P}(\limsup A_n) \geq \limsup \mathbb{P}(A_n)$ (prove this!), so we obtain convergence in probability $\limsup \mathbb{P}(|X_n - X| > \varepsilon) \leq \mathbb{P}(|X_n - X| > \varepsilon \text{ i.o.}) = 0$.
2. Note that $\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(|X_n - X|^p > \varepsilon^p)$. Now use Chebyshev.
3. Use $\|X_n - X\|_q \leq \|X_n - X\|_p$.
4. f is bounded, i.e., $|f(x)| \leq K/2$ for some $K < \infty$, and continuous, i.e., for any $\varepsilon > 0$ there is a $\delta > 0$ such that $|f(x) - f(y)| > \varepsilon$ implies $|x - y| > \delta$. Note that

$$|\mathbb{E}(f(X_n) - f(X))| \leq \mathbb{E}(|f(X_n) - f(X)|) = \mathbb{E}((I_{|f(X_n)-f(X)|>\varepsilon} + I_{|f(X_n)-f(X)|\leq\varepsilon})\,|f(X_n) - f(X)|).$$

Now $\mathbb{E}(I_{|f(X_n)-f(X)|\leq\varepsilon}|f(X_n) - f(X)|) \leq \varepsilon$, while by boundedness of f we obtain $\mathbb{E}(I_{|f(X_n)-f(X)|>\varepsilon}|f(X_n) - f(X)|) \leq K\,\mathbb{E}(I_{|f(X_n)-f(X)|>\varepsilon})$. Thus

$$|\mathbb{E}(f(X_n) - f(X))| \leq \varepsilon + K\,\mathbb{P}(|f(X_n) - f(X)| > \varepsilon) \leq \varepsilon + K\,\mathbb{P}(|X_n - X| > \delta)$$

where we have used continuity of f. The rightmost term converges to zero as $X_n \to X$ in probability, so we find that $\limsup |\mathbb{E}(f(X_n) - f(X))| \leq \varepsilon$. But this holds for any $\varepsilon > 0$, so evidently $\mathbb{E}(f(X_n)) \to \mathbb{E}(f(X))$ and we are done. □

These are the only implications that hold *in general*. Though this proposition is very useful in practice, you will perhaps get the most intuition about these modes of convergence by thinking about the following counterexamples.

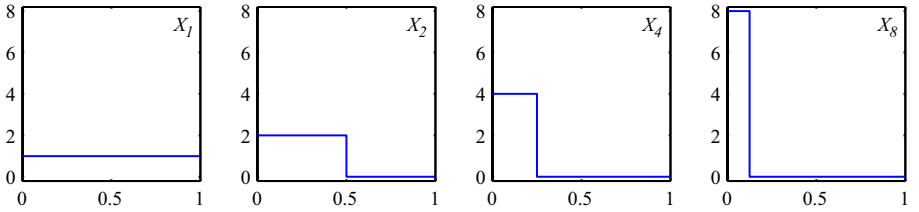

Figure 1.2. The random variables X_1, X_2, X_4, X_8 of example 1.5.4. The horizontal axis represents $\Omega = [0, 1]$, the vertical axis the value of $X_n(\omega)$. Note that the probability that X_n is nonzero shrinks to zero, but the value of X_n when it is nonzero becomes increasingly large.

Example 1.5.3 (Convergence in law but not in probability). It is easy to find counterexamples for this case; here is one of the simplest. Let $\Omega = \{a_1, a_2\}$, $\mathcal{F}$ is the power set, and $\mathbb{P}$ is the uniform measure ($\mathbb{P}(\{a_1\}) = 1/2$). Define the random variable $X(a_1) = 1$, $X(a_2) = -1$, and consider the sequence $X_n = (-1)^n X$. Obviously this sequence can never converge in probability, a.s., or in $\mathcal{L}^p$. However, $\mathbb{E}(f(X_n)) = \mathbb{E}(f(X))$ for any f, so $X_n \to X$ in law (and also $X_n \to -X$ in law!) Evidently this type of convergence has essentially no implication for the behavior of the random process X_n; certainly it does not look anything like convergence if we look at the paths $n \mapsto X_n(\omega)$ for fixed ω! (See figure 1.1). On the other hand, this is precisely the notion of convergence used in the central limit theorem.

The following three examples use the following probability space: $\Omega = [0, 1]$, $\mathcal{F} = \mathcal{B}([0, 1])$, and $\mathbb{P}$ is the uniform measure on $[0, 1]$ (under which $\mathbb{P}([a, b]) = b - a$).

Example 1.5.4 (Convergence a.s. but not in $\mathcal{L}^p$). Consider the sequence of random variables $X_n(\omega) = n\, I_{]0,1/n]}(\omega)$ (with $\omega \in \Omega = [0, 1]$). Then $X_n \to 0$ a.s.: for any ω, it is easy to see that $I_{]0,1/n]}(\omega) = 0$ for n sufficiently large. However, $\|X_n\|_1 = n\,\mathbb{E}(I_{]0,1/n]}) = 1$ for every n, so $X_n \not\to 0$ in $\mathcal{L}^1$. As convergence in $\mathcal{L}^p$ implies convergence in $\mathcal{L}^1$ (for $p \geq 1$), we see that X_n does not converge in $\mathcal{L}^p$. What is going on? Even though $\mathbb{P}(X_n \neq 0)$ shrinks to zero as $n \to \infty$, the value of X_n on those rare occasions that $X_n \neq 0$ grows so fast with n that we do not have convergence in $\mathcal{L}^p$, see figure 1.2 (compare with the intuition: a random variable that is zero with very high probability can still have a very large mean, if the outliers are sufficiently large). Note that as X_n converges a.s., it also converges in probability, so this example also shows that convergence in probability does not imply convergence in $\mathcal{L}^p$.

Example 1.5.5 (Convergence in $\mathcal{L}^q$ but not in $\mathcal{L}^p$). Let $X_n(\omega) = n^{1/p}\, I_{]0,1/n]}(\omega)$. You can easily verify that $X_n \to 0$ in $\mathcal{L}^q$ for all $q < p$, but not for $q \geq p$. Intuitively, $X_n \to X$ in $\mathcal{L}^q$ guarantees that the outliers of $|X_n - X|$ do not grow "too fast."

Example 1.5.6 (Convergence in $\mathcal{L}^p$ but not a.s.). This example is illustrated in figure 1.3; you might want to take a look at it first. Define X_n as follows. Write n as a binary number, i.e., $n = \sum_{i=0}^{\infty} n_i 2^i$ where $n_i \in \{0, 1\}$. Let k be the largest integer such that $n_k = 1$. Then we set $X_n(\omega) = I_{](n-2^k)2^{-k},(n-2^k+1)2^{-k}]}(\omega)$. It is not difficult to see

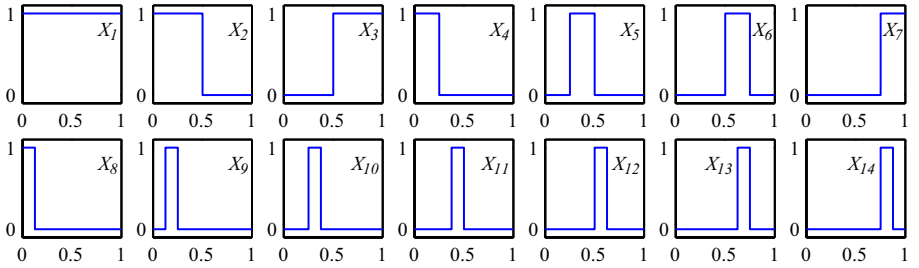

Figure 1.3. The random variables X_n, $n = 1, \ldots, 14$ of example 1.5.6. The horizontal axis represents $\Omega = [0, 1]$, the vertical axis the value of $X_n(\omega)$. Note that the probability that X_n is nonzero shrinks to zero, but $\limsup_{n\to\infty} X_n(\omega) = 1$ for any $\omega > 0$.

that $X_n \to 0$ in $\mathcal{L}^p$ for any p: indeed, $\mathbb{E}(|X_n|^p) = \mathbb{E}(X_n) = 2^{-k} \to 0$ as $n \to \infty$. However $X_n(\omega) \not\to 0$ for any $\omega > 0$; after all, for any n there is an $N(\omega) > n$ such that $X_{N(\omega)}(\omega) = 1$; hence $X_n(\omega) = 1$ infinitely often for every $\omega > 0$, and we most certainly do not have a.s. convergence. The occurence of $X_n(\omega) = 1$ becomes increasingly rare, however, when n is large, so that nonetheless the *probability* that $X_n > 0$ goes to zero (and as outliers are not an issue, the same holds in $\mathcal{L}^p$).

Before you move on, take some time to make sure you understand the various notions of convergence, their properties and their relations.

If you take a piece of paper, write on it all the modes of convergence which we have discussed, and draw arrows in both directions between each pair of convergence concepts, you will find that every one of these arrows is either implied by proposition 1.5.2 or ruled out, *in general*, by one of our counterexamples. However, if we impose some additional conditions then we can often still obtain some of the opposite implications (needless to say, our examples above will have to violate these conditions). An important related question is the following: if $X_n \to X$ in a certain sense, when does this imply that $\mathbb{E}(X_n) \to \mathbb{E}(X)$? The remainder of this section provides some answers to these questions. We will not strive for generality, but concentrate on the most widely used results (which we will have ample occasion to use).

Let us first tackle the question: when does convergence in probability imply a.s. convergence? To get some intuition, we revisit example 1.5.6.

Example 1.5.7. The following is a generalization of example 1.5.6. We construct a sequence of $\{0, 1\}$-valued random variables X_n such that $\mathbb{P}(X_n > 0) = \ell(n)$, where $\ell(n)$ is arbitrary (except that it is $[0, 1]$-valued). Set $X_1(\omega) = I_{]0,\ell(1)]}(\omega)$, then set $X_2(\omega) = I_{]\ell(1),\ell(1)+\ell(2)] \bmod [0,1]}(\omega)$, etc., so that each X_n is the indicator on the interval of length $\ell(n)$ immediately adjacent to the right of the interval corresponding to X_{n-1}, and we wrap around from 1 to 0 if necessary. Obviously $X_n \to 0$ in probability iff $\ell(n) \to 0$ as $n \to \infty$. We now ask: when does $X_n \to 0$ a.s.? If this is

not the case, then X_n must revisit every section of the interval $[0,1]$ infinitely many times; this means that the total "distance" travelled must be infinite $\sum_n \ell(n) = \infty$. On the other hand, if $\sum_n \ell(n) < \infty$, then eventually the right endpoint of the interval corresponding to X_n has to accumulate at some $x^* \in [0,1]$, so that $X_n \to 0$ a.s.

This example suggests that $X_n \to X$ in probability would imply $X_n \to X$ a.s. if only the convergence in probability happens "fast enough". This is generally true, and gives a nice application of the Borel-Cantelli lemma.

Lemma 1.5.8. *Let X_n, X be random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ s.t. $X_n \to X$ in probab. If $\sum_n \mathbb{P}(|X_n - X| > \varepsilon) < \infty$ for any $\varepsilon > 0$, then $X_n \to X$ a.s.*

Proof. By the Borel-Cantelli lemma $\mathbb{P}(|X_n - X| > \varepsilon \text{ i.o.}) = 0$ for any $\varepsilon > 0$. Thus

$$\mathbb{P}\left(\text{for any } k \in \mathbb{N},\ |X_n - X| > \frac{1}{k} \text{ for a finite number of } n \text{ only}\right) = 1 \quad \text{(why?)}.$$

Now use the usual calculus definition of convergence of a sequence $X_n(\omega)$. □

The following corollary will sometimes be useful.

Corollary 1.5.9. *Suppose that $X_n \to X$ in probability as $n \to \infty$. Then there exists a subsequence $n(k) \nearrow \infty$ such that $X_{n(k)} \to X$ a.s. as $k \to \infty$.*

Proof. As $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$ for $\varepsilon > 0$, there is for any $\varepsilon > 0$ and $\delta > 0$ an N such that $\mathbb{P}(|X_n - X| > \varepsilon) < \delta$ for all $n \geq N$. Choose $n(1)$ such that $\mathbb{P}(|X_{n(1)} - X| > 1/2) < 1/2$; now let $n(k)$ be such that $\mathbb{P}(|X_{n(k)} - X| > 2^{-k}) < 2^{-k}$ and $n(k) > n(k-1)$. Then evidently the sequence $X_{n(k)}$ satisfies the condition of the previous result. □

Our remaining goal is to find conditions under which convergence in probability implies convergence in $\mathcal{L}^p$. Before we can make progress in this direction, we need to introduce two fundamental results about convergence of expectations which are by themselves extremely useful and widely used.

Theorem 1.5.10 (Monotone convergence). *Let $\{X_n\}$ be a sequence of random variables such that $0 \leq X_1 \leq X_2 \leq \ldots$ a.s. Then $\mathbb{E}(X_n) \nearrow \mathbb{E}(\lim_{n\to\infty} X_n)$.*

For sequences X_n of *simple* functions the theorem is trivial: this is just the definition of the expectation! It only remains to extend the statement to general sequences. Note that the theorem holds even if $\mathbb{E}(\lim_{n\to\infty} X_n) = \infty$.

Proof. We can assume that $\{X_n(\omega)\}$ is nondecreasing for all $\omega \in \Omega$ by setting $X_n(\omega) = 0$ for those ω where this is not the case; this will not change the expectations (and hence the statement of the theorem), as by assumption the set of these ω has zero probability.

For every X_n, let $\{X_n^k\}_{k\in\mathbb{N}}$ be the approximating sequence of simple functions as in lemma 1.3.12, and let $\{X^k\}$ be the approximating sequence for $X = \lim_{n\to\infty} X_n$ (the latter exists by monotonicity of X_n, though it may take the value ∞). By construction $X_n^k \nearrow X_n$, $X^k \nearrow X$ as $k \to \infty$, and you can verify directly that $X_n^k \nearrow X^k$ as $n \to \infty$. By the definition of the expectation $\lim_{k\to\infty} \lim_{n\to\infty} \mathbb{E}(X_n^k) = \mathbb{E}(X)$. But $X_n^k \leq X_n$ (as X_n^k is nondecreasing), so $\mathbb{E}(X) \leq \lim \mathbb{E}(X_n)$. On the other hand $X_n \leq X$, so $\lim \mathbb{E}(X_n) \leq \mathbb{E}(X)$. □

Lemma 1.5.11 (Fatou's lemma). *Let $\{X_n\}$ be a sequence of a.s. nonnegative random variables. Then $\mathbb{E}(\liminf X_n) \le \liminf \mathbb{E}(X_n)$. In there exists a Y such that $X_n \le Y$ a.s. for all n and $\mathbb{E}(Y) < \infty$, then $\mathbb{E}(\limsup X_n) \ge \limsup \mathbb{E}(X_n)$.*

This should really be a theorem, but as it is called "Fatou's lemma" we will conform. The second half of the result is sometimes called the "reverse Fatou's lemma".

Proof. Define $Z_n = \inf_{k \ge n} X_k$, so $\liminf_n X_n = \lim_n Z_n$. Note that Z_n is nondecreasing, so by monotone convergence $\mathbb{E}(\liminf_n X_n) = \lim_n \mathbb{E}(Z_n)$. But $\mathbb{E}(Z_n) \le \inf_{k \ge n} \mathbb{E}(X_k)$ (why?), so $\mathbb{E}(\liminf_n X_n) \le \liminf_n \mathbb{E}(X_n)$. The second half of the result follows by applying the first half to the sequence $X'_n = Y - X_n$. □

We can now proceed to find a useful condition when convergence in probability implies convergence in $\mathcal{L}^p$. What sort of condition can we expect? Recall that intuitively, a sequence $X_n \to X$ in probability should converge in $\mathcal{L}^p$ if the outliers of $|X_n - X|$ do not grow too fast. A good way to control the outliers is to impose suitable boundedness conditions on $|X_n - X|$, which is precisely what we will do.

Theorem 1.5.12 (Dominated convergence). *Let $X_n \to X$ in probability, and suppose there exists a nonnegative $Y \in \mathcal{L}^p$, with $p \ge 1$, such that $|X_n| \le Y$ a.s. for all n. Then X and X_n are in $\mathcal{L}^p$ for all n, $X_n \to X$ in $\mathcal{L}^p$, and $\mathbb{E}(X_n) \to \mathbb{E}(X)$.*

Proof. We begin by proving the theorem assuming that $X_n \to X$ a.s. At the end of the day, we will weaken this to convergence in probability.

First, note that $|X_n| \le Y$ implies that $\mathbb{E}(|X_n|^p) \le \mathbb{E}(Y^p) < \infty$, so $X_n \in \mathcal{L}^p$ for all n. As $X_n \to X$ a.s., $|X| \le Y$ as well, so $X \in \mathcal{L}^p$. Next, note that $|X_n - X|^p \le (|X_n| + |X|)^p \le (2Y)^p$, and the latter is integrable by assumption. Hence by Fatou's lemma $\limsup \mathbb{E}(|X_n - X|^p) \le \mathbb{E}(\limsup |X_n - X|^p) = 0$, so $X_n \to X$ in $\mathcal{L}^p$. But convergence in $\mathcal{L}^p$ ($p \ge 1$) implies convergence in $\mathcal{L}^1$, so that $|\mathbb{E}(X_n) - \mathbb{E}(X)| \le \mathbb{E}(|X_n - X|) \to 0$.

Now suppose that $X_n \to X$ in probability (rather than a.s.), but $\|X_n - X\|_p$ does not converge to zero. Then there is a subsequence $n(k) \nearrow \infty$ such that $\|X_{n(k)} - X\|_p \to \varepsilon$ for some $\varepsilon > 0$. But clearly $X_{n(k)} \to X$ in probability, so by corollary 1.5.9 there exists a further subsequence $n'(k)$ such that $X_{n'(k)} \to X$ a.s. But then by the a.s. version of dominated convergence $\|X_{n'(k)} - X\|_p \to 0$ as $k \to \infty$, which contradicts $\|X_{n'(k)} - X\|_p \to \varepsilon$. □

A special case of the dominated convergence theorem is used particularly often: if the sequence $\{X_n\}$ is uniformly bounded, i.e., there is some $K < \infty$ such that $|X_n| \le K$ a.s. for all n, then $X_n \to X$ in probability (or a.s.) gives $\mathbb{E}(X_n) \to \mathbb{E}(X)$.

Let us finally discuss one convergence result of a somewhat different nature. Note that all the convergence theorems above assume that we already know that our sequence converges to a particular random variable $X_n \to X$; they only allow us to convert between one mode of convergence and another. However, we are often just given a sequence X_n, and we still need to establish that X_n converges to *something*. The following method allows us to establish that a sequence X_n has a limit, without having to know in advance what that limit is. We will encounter another way to show that a sequence converges in the next chapter (the *martingale convergence theorem*).

Definition 1.5.13. A sequence $\{X_n\}$ of random variables in $\mathcal{L}^p$, $p \ge 1$ is called a *Cauchy sequence (in $\mathcal{L}^p$)* if $\sup_{m,n \ge N} \|X_m - X_n\|_p \to 0$ as $N \to \infty$.

Proposition 1.5.14 (Completeness of $\mathcal{L}^p$). *Let X_n be a Cauchy sequence in $\mathcal{L}^p$. Then there exists a random variable $X_\infty \in \mathcal{L}^p$ such that $X_n \to X_\infty$ in $\mathcal{L}^p$.*

Remark 1.5.15. As you know from your calculus course, $\mathbb{R}^n$ also has the property that any Cauchy sequence converges: if $\sup_{m,n\geq N} |x_m - x_n| \to 0$ as $N \to \infty$ for a sequence $\{x_n\} \subset \mathbb{R}^n$, then there is an $x_\infty \in \mathbb{R}^n$ such that $x_n \to x_\infty$. In fact, many (but not all) metric spaces have this property, so it is not shocking that it is true also for $\mathcal{L}^p$. A metric space in which every Cauchy sequence converges is called *complete*.

Proof of proposition 1.5.14. We need to do two things: first, we need to identify a candidate X_∞. Once we have constructed such an X_∞, we need to show that $X_n \to X_\infty$ in $\mathcal{L}^p$.

Let $M(N) \nearrow \infty$ be a subsequence such that $\sup_{m,n\geq M(N)} \|X_n - X_m\|_p \leq 2^{-N}$ for all N. As $\|\cdot\|_1 \leq \|\cdot\|_p$ (recall that we assume $p \geq 1$), this implies $\sup_{m,n\geq M(N)} \mathbb{E}(|X_n - X_m|) \leq 2^{-N}$, and in particular $\mathbb{E}(|X_{M(N+1)} - X_{M(N)}|) \leq 2^{-N}$. Hence

$$\mathbb{E}\left(\sum_{n=1}^{\infty} |X_{M(n+1)} - X_{M(n)}|\right) = \sum_{n=1}^{\infty} \mathbb{E}\left(|X_{M(n+1)} - X_{M(n)}|\right) < \infty,$$

where we have used the monotone convergence theorem to exchange the summation and the expectation. But then the series $X_{M(n)} = X_{M(1)} + \sum_{k=2}^{n}(X_{M(k)} - X_{M(k-1)})$ is a.s. absolutely convergent, so $X_{M(n)}$ converges a.s. to some random variable X_∞. Moreover,

$$\mathbb{E}(|X_{M(k)} - X_\infty|^p) = \mathbb{E}\left(\liminf_{n\to\infty} |X_{M(k)} - X_n|^p\right) \leq \liminf_{n\to\infty} \mathbb{E}(|X_{M(k)} - X_n|^p) \leq 2^{-kp}$$

using Fatou's lemma, so we conclude that $X_{M(k)} \to X_\infty$ in $\mathcal{L}^p$, and in particular $X_\infty \in \mathcal{L}^p$ itself (the latter follows as evidently $X_{M(k)} - X_\infty \in \mathcal{L}^p$, $X_{M(k)} \in \mathcal{L}^p$ by assumption, and $\mathcal{L}^p$ is linear, so $X_\infty = X_{M(k)} - (X_{M(k)} - X_\infty) \in \mathcal{L}^p$).

It remains to show that $X_n \to X_\infty$ in $\mathcal{L}^p$ (i.e., not necessarily for the subsequence $M(n)$). To this end, note that $\|X_n - X_\infty\|_p \leq \|X_n - X_{M(n)}\|_p + \|X_{M(n)} - X_\infty\|_p$; that the second term converges to zero we have already seen, while that the first term converges to zero follows directly from the fact that X_n is a Cauchy sequence. Thus we are done. □

1.6 Induced measures, independence, and absolute continuity

We have seen that the construction of a useful probability space is not a trivial task. It was, perhaps surprisingly, not straightforward to define a σ-algebra that is sufficiently small to be useful; and constructing a suitable probability measure on such a σ-algebra is something we swept under the carpet even in one of the simplest cases (theorem 1.1.15). Fortunately we will not need to appeal to the precise details of these constructions; once a probability space has been constructed, it is fairly easy to use.

Up to this point, however, we have not constructed any probability space that is more complicated than $\Omega = \mathbb{R}$. The theme of this section is: given an existing probability space $(\Omega, \mathcal{F}, \mathbb{P})$, how can we transform this space into a different (and potentially more complicated or interesting) probability space? The techniques of this section will be sufficient to last us throughout this course.

Induced measures and laws

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $(S, \mathcal{S})$ be another space with a suitable σ-algebra $\mathcal{S}$. We want to construct a probability measure μ on $\mathcal{S}$, but we do not wish to go about this from scratch (having gone through all the trouble to construct $\mathbb{P}$ already!) Fortunately there is an easy way to "inherit" μ from $\mathbb{P}$.

Definition 1.6.1. If $h : \Omega \to S$ is measurable, then the map $\mu : \mathcal{S} \to [0, 1]$ defined by $\mu(A) = \mathbb{P}(h^{-1}(A))$ is a probability measure (why?), called the *induced measure*.

If we interpret the map h as an S-valued random variable, then the induced measure μ_h on S is called the *law of* h or the *distribution of* h. This familiar concept is often used to characterize random variables, for example, a (scalar) Gaussian random variable is a random variable X whose law μ_X is a Gaussian measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ (in the sense of example 1.1.16). Given the law of X, the probability of observing $X \in A$ (for any $A \in \mathcal{S}$) can always be calculated as $\mu_X(A)$.

Independence and product spaces

Let us begin by recalling the basic notions of *independence*. This should already be very familiar to you, certainly on an intuitive level.

Definition 1.6.2. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

1. A countable set of events $\{A_n\}_{n\in I} \subset \mathcal{F}$ are *independent* if
$$\mathbb{P}(A_{k_1} \cap A_{k_2} \cap \cdots \cap A_{k_n}) = \mathbb{P}(A_{k_1})\,\mathbb{P}(A_{k_2}) \cdots \mathbb{P}(A_{k_n}),$$
for all finite ($n < \infty$) subsets $\{k_1, \ldots, k_n\}$ of the index set I.
2. A countable set $\{\mathcal{G}_n\}_{n\in I}$ of σ-algebras $\mathcal{G}_n \subset \mathcal{F}$ are *independent* if any finite set of events $A_1, \ldots, A_n$ from distinct $\mathcal{G}_k$ are independent.
3. A countable set $\{X_n\}_{n\in I}$ of random variables are *independent* if the σ-algebras generated by these random variables $\mathcal{X}_n = \sigma(X_n)$ are independent.

Example 1.6.3. We throw two dice. Thus $\Omega = \{(1,1), (1,2), \ldots, (6,6)\}$, $\mathcal{F}$ is the power set and $\mathbb{P}(\{(i,j)\}) = 1/36$ for all $(i,j) \in \Omega$. Define the random variables $X_1((i,j)) = i$ and $X_2((i,j)) = j$, corresponding to the outcome of the first and second die, respectively. Then $\mathbb{P}(X_1 \in A \text{ and } X_2 \in B) = \mathbb{E}(I_A(X_1)I_B(X_2)) = \sum_{(i,j)} \mathbb{P}(\{(i,j)\})I_A(i)I_B(j) = \frac{1}{6}\#A \times \frac{1}{6}\#B = \mathbb{P}(X_1 \in A)\,\mathbb{P}(X_2 \in B)$ for any sets $A, B \subset \{1, \ldots, 6\}$. But $\sigma(X_1)$ consists of $\varnothing$ and all sets of the form $X_1^{-1}(A)$, and similarly for $\sigma(X_2)$. Hence X_1 and X_2 are independent.

Example 1.6.4. The last example suggests a way to construct probability spaces that carry independent events. Suppose we start with two *discrete* probability spaces $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$, where $\mathcal{F}_1$ and $\mathcal{F}_2$ are the power sets. Now consider the space $(\Omega, \mathcal{F}, \mathbb{P})$ where $\Omega = \Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) : \omega_1 \in \Omega_1,\ \omega_2 \in \Omega_2\}$, $\mathcal{F}$ is the power set, and $\mathbb{P}(\{(\omega_1, \omega_2)\}) = \mathbb{P}_1(\{\omega_1\})\,\mathbb{P}_2(\{\omega_2\})$. The individual probability spaces are naturally embedded in $(\Omega, \mathcal{F}, \mathbb{P})$: if we define the projection maps

$\rho_1 : (\omega_1, \omega_2) \mapsto \omega_1$ and $\rho_2 : (\omega_1, \omega_2) \mapsto \omega_2$, then $\mathbb{P}(\rho_1^{-1}(A)) = \mathbb{P}_1(A)$ and $\mathbb{P}(\rho_2^{-1}(B)) = \mathbb{P}_2(B)$ for any $A \in \mathcal{F}_1$, $B \in \mathcal{F}_2$. You can now easily repeat the previous example to conclude that $\rho_1^{-1}(\mathcal{F}_1)$ and $\rho_2^{-1}(\mathcal{F}_2)$ are independent under $\mathbb{P}$.

We have just shown off the notion of a *product space*. A very similar idea holds in the general case, except that there will be some unpleasantness in proving that the product measure $\mathbb{P} = \mathbb{P}_1 \times \mathbb{P}_2$ is well defined (actually this is quite straightforward; see, e.g., [Wil91, chapter 8] or [Bil86, sec. 18]). Let us just state the facts.

Definition 1.6.5. Given two measurable spaces $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$, we define the *product space* $\Omega_1 \times \Omega_2 \equiv \{(\omega_1, \omega_2) : \omega_1 \in \Omega_1,\ \omega_2 \in \Omega_2\}$. The *product σ-algebra* on $\Omega_1 \times \Omega_2$ is defined as $\mathcal{F}_1 \times \mathcal{F}_2 \equiv \sigma\{\rho_1, \rho_2\}$, where $\rho_i : \Omega_1 \times \Omega_2 \to \Omega_i$ are the projection maps $\rho_1 : (\omega_1, \omega_2) \mapsto \omega_1$ and $\rho_2 : (\omega_1, \omega_2) \mapsto \omega_2$.

Theorem 1.6.6. *Let $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ be two probability spaces, and denote by $\rho_1 : \Omega_1 \times \Omega_2 \to \Omega_1$ and $\rho_2 : \Omega_1 \times \Omega_2 \to \Omega_2$ the projection maps.*

1. *There exists a unique probability measure $\mathbb{P}_1 \times \mathbb{P}_2$ on $\mathcal{F}_1 \times \mathcal{F}_2$, called the product measure, under which the law of ρ_1 is $\mathbb{P}_1$, the law of ρ_2 is $\mathbb{P}_2$, and ρ_1 and ρ_2 are independent Ω_1- and Ω_2-valued random variables, respectively.*

2. *The construction is in some sense unique: suppose that on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ are defined an S-valued random variable X (with law μ_X) and a T-valued random variable Y (with law μ_Y). Then the law of the $S \times T$-valued random variable (X, Y) is the product measure $\mu_X \times \mu_Y$.*

3. *If $f : \Omega_1 \times \Omega_2 \to \mathbb{R}$ is $\mathcal{F}_1 \times \mathcal{F}_2$-measurable, then $f(\omega_1, \cdot) : \Omega_2 \to \mathbb{R}$ is $\mathcal{F}_2$-measurable for all ω_1, and $f(\cdot, \omega_2) : \Omega_1 \to \mathbb{R}$ is $\mathcal{F}_1$-measurable for all ω_2.*

4. ***(Tonelli)** If $f : \Omega_1 \times \Omega_2 \to \mathbb{R}$ is $\mathcal{F}_1 \times \mathcal{F}_2$-measurable and $f \geq 0$ a.s., then*

$$\mathbb{E}_{\mathbb{P}_1 \times \mathbb{P}_2}(f) = \mathbb{E}_{\mathbb{P}_2}\left[\int f(\omega_1, \omega_2)\, \mathbb{P}_1(d\omega_1)\right] = \mathbb{E}_{\mathbb{P}_1}\left[\int f(\omega_1, \omega_2)\, \mathbb{P}_2(d\omega_2)\right].$$

   In particular, these expressions make sense (the inner integrals are measurable).

5. ***(Fubini)** The previous statement still holds for random variables f that are not necessarily nonnegative, provided that $\mathbb{E}(|f|) < \infty$.*

The construction extends readily to products of a finite number of spaces. Starting from, e.g., the simple probability space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_N)$ where $\mathbb{P}_N$ is a Gaussian measure (with mean zero and unit variance, say), which we have already constructed in example 1.1.16, we can now construct a larger space $(\Omega, \mathcal{F}, \mathbb{P})$ that carries a finite number d of independent copies of a Gaussian random variable: set $\Omega = \mathbb{R} \times \cdots \times \mathbb{R}$ (d times), $\mathcal{F} = \mathcal{B}(\mathbb{R}) \times \cdots \times \mathcal{B}(\mathbb{R})$, and $\mathbb{P} = \mathbb{P}_N \times \cdots \times \mathbb{P}_N$. The independent Gaussian random variables are then precisely the projection maps $\rho_1, \ldots, \rho_d$.

Remark 1.6.7. The Fubini and Tonelli theorems tell us that taking the expectation with respect to the product measure can often be done more simply: if we consider the product space random variable $f : \Omega_1 \times \Omega_2 \to \mathbb{R}$ as an ω_1-valued random variable

$f(\omega_1, \cdot) : \Omega_2 \to \mathbb{R}$, then we may take the expectation of this random variable with respect to $\mathbb{P}_2$. The resulting map can subsequently be interpreted as an ω_1-valued random variable, whose $\mathbb{P}_1$-expectation we can calculate. The Fubini and Tonelli theorems tell us that (under mild regularity conditions) it does not matter whether we apply $\mathbb{E}_{\mathbb{P}_1 \times \mathbb{P}_2}$, or $\mathbb{E}_{\mathbb{P}_1}$ first and then $\mathbb{E}_{\mathbb{P}_2}$, or vice versa.

We will more often encounter these results in a slightly different context. Suppose we have a continuous time stochastic process, i.e., a collection of random variables $\{X_t\}_{t\in[0,T]}$ (see section 2.4 for more on this concept). Such a process is called *measurable* if the map $X_\cdot : [0,T] \times \Omega \to \mathbb{R}$ is $\mathcal{B}([0,T]) \times \mathcal{F}$-measurable. In this case,

$$Y(\omega) = \int_0^T X_t(\omega)\, dt$$

is a well defined random variable: you can interpret this as T times the expectation of $X_\cdot(\omega) : [0,T] \to \mathbb{R}$ with respect to the uniform measure on $[0,T]$ (this is the *Lebesgue measure* of remark 1.3.11, restricted to $[0,T]$). Suppose that we are interested in the expectation of Y; it is then often useful to know whether we can exchange the order of integration and expectation, i.e., whether

$$\mathbb{E}(Y) = \mathbb{E}\left(\int_0^T X_t\, dt\right) = \int_0^T \mathbb{E}(X_t)\, dt.$$

The Fubini and Tonelli theorems give sufficient conditions for this to be the case.

As you can tell from the preceding remark, product spaces play an important role even in cases that have nothing to do with independence. Let us get back to the theme of this section, however, which was to build more complicated probability spaces from simpler ones. We have seen how to build a product probability space with a finite number of independent random variables. In our construction of the Wiener process, however, we will need an entire sequence of independent random variables. The construction of the product space and σ-algebra extends trivially to this case (how?), but the construction of an infinite product measure brings with it some additional difficulties. Nonetheless this can always be done [Kal97, corollary 5.18]. For the purposes of this course, however, the following theorem is all we will need.

Theorem 1.6.8. *Let $\{\mathbb{P}_n\}$ be a sequence of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Then there exists a probability measure $\mathbb{P}$ on $(\mathbb{R}\times\mathbb{R}\times\cdots, \mathcal{B}(\mathbb{R})\times\mathcal{B}(\mathbb{R})\times\cdots)$ such that the projection maps $\rho_1, \rho_2, \ldots$ are independent and have the law $\mathbb{P}_1, \mathbb{P}_2, \ldots$, respectively.*

The proof of this result is so much fun that it would be a shame not to include it. However, the method of proof is quite peculiar and will not really help you in the rest of this course. Feel free to skip it completely, unless you are curious.

Rather than construct the infinite product measure directly, the idea of the proof is to construct a sequence of independent random variables $\{X_n\}$ on the (surprisingly simple!) probability space $([0,1], \mathcal{B}([0,1]), \lambda)$, where λ is the uniform measure, whose laws are $\mathbb{P}_1, \mathbb{P}_2, \ldots$, respectively. The theorem then follows trivially, as the law of the $\mathbb{R}\times\mathbb{R}\times\cdots$-valued random variable $X = (X_1, X_2, \ldots)$ is then precisely the desired product measure $\mathbb{P}$.

It may seem a little strange that we can construct all these independent random variables on such a simple space as $([0,1], \mathcal{B}([0,1]), \lambda)$! After all, this would be the natural space on which to construct a single random variable uniformly distributed on the interval. Strange things are possible, however, because $[0,1]$ is continuous—we can cram a lot of information in there, if we encode it correctly. The construction below works precisely in this way. We proceed in two steps. First, we show that we can dissect and then reassemble the interval in such a way that it gives an entire sequence of random variables uniformly distributed on the interval. It then remains to find functions of these random variables that have the correct laws $\mathbb{P}_1, \mathbb{P}_2, \ldots$.

The dissection and reassembly of the interval $[0,1]$ is based on the following lemma.

Lemma. *Let ξ be a random variable that is uniformly distributed on $[0,1]$, and denote by $\xi = \sum_{n=1}^{\infty} \xi_n 2^{-n}$ its binary expansion (i.e., ξ_n are $\{0,1\}$-valued random variables). Then all the ξ_n are independent and take the values 0 and 1 with equal probability. Conversely, if $\{\xi_n\}$ is any such sequence, then $\xi = \sum_{n=1}^{\infty} \xi_n 2^{-n}$ is uniformly distributed on $[0,1]$.*

Proof. Consider the first k binary digits ξ_n, $n \leq k$. If you write down all possible combinations of k zeros and ones, and partition $[0,1]$ into sets whose first k digits coincide to each of these combinations, you will find that $[0,1]$ is partitioned into 2^k equally sized sets, each of which has probability 2^{-k}. As for every $n \leq k$ the set $\{\xi_n = 1\}$ is the union of 2^{-k+1} sets in our partition, we find that every such set has probability 2^{-1}. But then clearly the ξ_n, $n \leq k$ must be independent (why?) As this holds for any $k < \infty$, the first part of the lemma follows. The second part of the lemma follows directly from the first part, as any random variable constructed in this way must have the same law as the random variable ξ considered in the first part. □

Corollary. *There exists a sequence $\{Y_n\}$ of independent random variables on the probability space $([0,1], \mathcal{B}([0,1]), \lambda)$, each of which is uniformly distributed on the unit interval $[0,1]$.*

Proof. Define $\xi : [0,1] \to \mathbb{R}$, $\xi(x) = x$. Then ξ is uniformly distributed on $[0,1]$, so by the previous lemma its sequence of binary digits $\{\xi_n\}$ are independent and take the values $\{0,1\}$ with equal probability. Let us now reorder $\{\xi_n\}_{n\in\mathbb{N}}$ into a two-dimensional array $\{\tilde{\xi}_{mn}\}_{m,n\in\mathbb{N}}$ (i.e., each $\tilde{\xi}_{mn}$ coincides with precisely one ξ_n). This is easily done, for example, as in the usual proof that the rational numbers are countable. Define $Y_n = \sum_{m=1}^{\infty} \tilde{\xi}_{mn} 2^{-m}$. By the previous lemma, the sequence $\{Y_n\}$ has the desired properties. □

Proof of theorem 1.6.8. We have constructed a sequence $\{Y_n\}$ of independent uniformly distributed random variables on $[0,1]$. The last step of the proof consists of finding a sequence of measurable functions $f_n : [0,1] \to \mathbb{R}$ such that the law of $X_n = f_n(Y_n)$ is $\mathbb{P}_n$. Then we are done, as $\{X_n\}$ is a sequence of independent random variables with law $\mathbb{P}_1, \mathbb{P}_2, \ldots$, and the product measure $\mathbb{P}$ is then simply the law of $(X_1, X_2, \ldots)$ as discussed above.

To construct the functions f_n, let $F_n(x) = \mathbb{P}_n(]-\infty, x])$ be the CDF of the measure $\mathbb{P}_n$ (see theorem 1.1.15). Note thst F_n takes values in the interval $[0,1]$. Now define $f_n(u) = \inf\{x \in \mathbb{R} : u \leq F_n(x)\}$. Then $\lambda(f_n(Y_n) \leq y) = \lambda(Y_n \leq F_n(y)) = F_n(y)$, as Y_n is uniformly distributed. Hence $f_n(Y_n)$ has the law $\mathbb{P}_n$, and we are done. □

Absolutely continuous measures and the Radon-Nikodym theorem

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a given probability space. It is often interesting to try to find other measures on $\mathcal{F}$ with different properties. We may have gone through some trouble to construct a measure $\mathbb{P}$, but once we have such a measure, we can generate a large family of related measures using a rather simple technique. This idea will come in *very* handy in many situations; calculations which are difficult under one measure can

often become very simple if we change to a suitably modified measure (for example, if $\{X_n\}$ is a collection of random variables with some complicated dependencies under $\mathbb{P}$, it may be advantageous to compute using a modified measure $\mathbb{Q}$ under which the X_n are independent). Later on, the change of measure concept will form the basis for one of the most basic tools in our stochastic toolbox, the Girsanov theorem.

The basic idea is as follows. Let f be a nonnegative random variable with unit expectation $\mathbb{E}(f) = 1$. For any set $A \in \mathcal{F}$, define the quantity

$$\mathbb{Q}(A) = \mathbb{E}_{\mathbb{P}}(I_A f) \equiv \int_A f(\omega)\, \mathbb{P}(d\omega).$$

Then $\mathbb{Q}(A)$ is itself a probability measure (why?), and moreover

$$\mathbb{E}_{\mathbb{Q}}(g) = \int g(\omega)\, \mathbb{Q}(d\omega) = \int g(\omega)\, f(\omega)\, \mathbb{P}(d\omega) = \mathbb{E}_{\mathbb{P}}(gf)$$

for any random variable g for which either side is well defined (why?).

Definition 1.6.9. A probability measure $\mathbb{Q}$ is said to have a *density* with respect to a probability measure $\mathbb{P}$ if there exists a nonnegative random variable f such that $\mathbb{Q}(A) = \mathbb{E}_{\mathbb{P}}(I_A f)$ for every measurable set A. The density f is denoted as $d\mathbb{Q}/d\mathbb{P}$.

Remark 1.6.10. In your introductory probability course, you likely encountered this idea very frequently with a minor difference: the concept still works if $\mathbb{P}$ is not a probability measure but, e.g., the Lebesgue measure of remark 1.3.11. We then define

$$\mathbb{Q}(A) = \int_A f(x)\, dx, \qquad \text{e.g.,} \qquad \mathbb{Q}([a,b]) = \int_a^b f(x)\, dx,$$

where f is now the density of $\mathbb{Q}$ *with respect to the Lebesgue measure*. Not all probability measures on $\mathbb{R}$ admit such a representation (consider example 1.1.13), but many interesting examples can be constructed, including the Gaussian measure (where $f \propto \exp(-(x-\mu)^2/2\sigma)$). Such expressions allow nice explicit computations in the case where the underlying probability space is $\mathbb{R}^d$, which is the reason that most introductory courses are centered around such objects (rather than introducing measure theory, which is needed for more complicated probability spaces).

Suppose that $\mathbb{Q}$ has a density f with respect to $\mathbb{P}$. Then these measures must satisfy an important consistency condition: if $\mathbb{P}(A) = 0$ for some event A, then $\mathbb{Q}(A)$ must also be zero. To see this, note that $I_A(\omega) f(\omega) = 0$ for $\omega \in A^c$ and $\mathbb{P}(A^c) = 1$, so $I_A f = 0$ $\mathbb{P}$-a.s. In other words, if $\mathbb{Q}$ has a density with respect to $\mathbb{P}$, then any event that never occurs under $\mathbb{P}$ certainly never occurs under $\mathbb{Q}$. Similarly, any event that happens with probability one under $\mathbb{P}$ must happen with probability one under $\mathbb{Q}$ (why?). Evidently, the use of a density to transform a probability measure $\mathbb{P}$ into another probability measure $\mathbb{Q}$ "respects" those events that happen for sure or never happen at all. This intuitive notion is formalized by the following concept.

Definition 1.6.11. A measure $\mathbb{Q}$ is said to be *absolutely continuous* with respect to a measure $\mathbb{P}$, denoted as $\mathbb{Q} \ll \mathbb{P}$, if $\mathbb{Q}(A) = 0$ for all events A such that $\mathbb{P}(A) = 0$.

We have seen that if $\mathbb{Q}$ has a density with respect to $\mathbb{P}$, then $\mathbb{Q} \ll \mathbb{P}$. It turns out that the converse is also true: if $\mathbb{Q} \ll \mathbb{P}$, then we can always find some density f such that $\mathbb{Q}(A) = \mathbb{E}_{\mathbb{P}}(I_A f)$. *Hence the existence of a density is completely equivalent to absolute continuity of the measures*. This is a deep result, known as the *Radon-Nikodym theorem*. It also sheds considerable light on the concept of a density: the intuitive meaning of the existence of a density is not immediately obvious, but the conceptual idea behind absolute continuity is clear.

Theorem 1.6.12 (Radon-Nikodym). *Suppose that $\mathbb{Q} \ll \mathbb{P}$ are two probability measures on the space $(\Omega, \mathcal{F})$. Then there exists a nonnegative $\mathcal{F}$-measurable function f with $\mathbb{E}_{\mathbb{P}}(f) = 1$, such that $\mathbb{Q}(A) = \mathbb{E}_{\mathbb{P}}(I_A f)$ for every $A \in \mathcal{F}$. Moreover, f is unique in the sense that if f' is another $\mathcal{F}$-measurable function with this property, then $f' = f$ $\mathbb{P}$-a.s. Hence it makes sense to speak of 'the' density, or Radon-Nikodym derivative, of $\mathbb{Q}$ with respect to $\mathbb{P}$, and this density is denoted as $d\mathbb{Q}/d\mathbb{P}$.*

In the case that Ω is a finite set, this result is trivial to prove. You should do this now: convince yourself that the equivalence between absolute continuity and the existence of a density is to be expected. The uniqueness part of the theorem also follows easily (why?), but the existence part is not so trivial in the general case. The theorem is often proved using functional analytic tools, notably the Riesz representation theorem; see e.g. [GS96]. A more measure-theoretic proof can be found in [Bil86]. Most beautiful is the probabilistic proof using martingale theory, see [Wil91, sec. 14.13]. We will follow this approach to prove the Radon-Nikodym theorem in section 2.2, after we have developed some more of the necessary tools.

1.7 A technical tool: Dynkin's π-system lemma

When developing the basic tools of probability there is an elementary technical result, known as Dynkin's π-system lemma, which is very often employed to complete certain proofs. Once the foundations have been laid and it comes to actually using the theory, we will no longer have much use for this technique (unlike, for example, the dominated convergence theorem, which we will use constantly); as such, we have avoided introducing this tool up to this point (though it has already secretly been used: for example, theorem 1.6.6 relies on this sort of reasoning!) Nonetheless we will need the π-system lemma briefly in chapters 2 and 3, and it is good to know that it exists in any case, so we will discuss the method briefly in this section.

The basic problem is as follows. Suppose we have defined, in one way or another, two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on some space $(\Omega, \mathcal{F})$. It could well be that $\mathbb{P} = \mathbb{Q}$, but as these measures were constructed in different ways this may not be so easy to prove. In particular, it would be rather tedious if we had to check $\mathbb{P}(A) = \mathbb{Q}(A)$ for every single set $A \in \mathcal{F}$. Dynkin's π-system lemma gives us a way to check the equality of two measures in a simpler way: roughly speaking, if we have verified $\mathbb{P}(A) = \mathbb{Q}(A)$ for "enough" sets $A \in \mathcal{F}$, then the measures must be equivalent.

Definition 1.7.1. Let Ω be a set. A *π-system* $\mathcal{C}$ is a collection of subsets of Ω that is closed under finite intersections, i.e., $A, B \in \mathcal{C}$ implies $A \cap B \in \mathcal{C}$. A *λ-system* $\mathcal{D}$ is a

collection of subsets of Ω such that $\Omega \in \mathcal{D}$, $A, B \in \mathcal{D}$ with $A \subset B$ implies $B \backslash A \in \mathcal{D}$, and for any sequence $\{A_n\}$ in $\mathcal{D}$ with $A_1 \subset A_2 \subset \cdots$, we have $\bigcup_n A_n \in \mathcal{D}$.

One could define a σ-algebra as a collection of subsets of Ω that is both a π-system and a λ-system (why?). Hence the following result should not come as a shock.

Lemma 1.7.2. *Let $\mathcal{C}$ be a π-system on some set Ω, and let $\mathcal{D} = \lambda\{\mathcal{C}\}$ be the smallest λ-system on Ω such that $\mathcal{C} \subset \mathcal{D}$. Then $\mathcal{D} = \sigma\{\mathcal{C}\}$.*

Proof. Note that the smallest λ-system that contains a (not necessarily countable) collection of sets is well defined: this is simply the intersection of all such λ-systems. Hence we are done if we can show that $\mathcal{D}$ is itself a π-system. After all, in that case it is a σ-algbera that contains $\mathcal{C}$, so $\sigma\{\mathcal{C}\} \subset \mathcal{D}$; on the other hand, $\sigma\{\mathcal{C}\}$ is a λ-system that contains $\mathcal{C}$, so $\mathcal{D} \subset \sigma\{\mathcal{C}\}$.

It thus remains to show that $A, B \in \mathcal{D}$ implies $A \cap B \in \mathcal{D}$. We first claim that this is true for $A \in \mathcal{D}$ and $B \in \mathcal{C}$. To see this, simply note that for any fixed $B \in \mathcal{C}$ we have $\mathcal{D} \subset \{A \subset \Omega : A \cap B \in \mathcal{D}\}$, as the latter is clearly a λ-system containing $\mathcal{C}$. We can thus conclude that for any fixed $A \in \mathcal{D}$, the collection $\{B \subset \Omega : A \cap B \in \mathcal{D}\}$ contains $\mathcal{C}$. But this collection is again a λ-system, so must contain $\mathcal{D}$. Hence $\mathcal{D}$ is a π-system. □

Lemma 1.7.3 (Dynkin). *Let $(\Omega, \mathcal{F})$ be a measurable space, and let $\mathcal{C}$ be a π-system such that $\mathcal{F} = \sigma\{\mathcal{C}\}$. If two probability measures $\mathbb{P}$ and $\mathbb{Q}$ agree on $\mathcal{C}$, i.e., if we have $\mathbb{P}(A) = \mathbb{Q}(A)$ for all $A \in \mathcal{C}$, then $\mathbb{P}$ and $\mathbb{Q}$ are equal ($\mathbb{P}(A) = \mathbb{Q}(A)$ for all $A \in \mathcal{F}$).*

Proof. Define $\mathcal{D} = \{A \in \mathcal{F} : \mathbb{P}(A) = \mathbb{Q}(A)\}$. You can verify directly that $\mathcal{D}$ is a λ-system, and by assumption $\mathcal{C} \subset \mathcal{D}$. But then $\mathcal{D} = \mathcal{F}$ by the previous lemma. □

For sake of example, let us prove the uniqueness part of theorem 1.6.6. Recall that $\rho_i : \Omega_1 \times \Omega_2 \to \Omega_i$ are the projection maps, and $\mathcal{F}_1 \times \mathcal{F}_2 = \sigma\{\rho_1, \rho_2\}$. Hence $\mathcal{F}_1 \times \mathcal{F}_2 = \sigma\{\mathcal{C}\}$ with $\mathcal{C} = \{A \times B : A \in \mathcal{F}_1,\ B \in \mathcal{F}_2\}$, which is clearly a π-system. Now any measure $\mathbb{P}$ under which ρ_i has law $\mathbb{P}_i$ and under which ρ_1 and ρ_2 are independent must satisfy $\mathbb{P}(A \times B) = \mathbb{P}_1(A)\,\mathbb{P}_2(B)$: this follows immediately from the definition of independence. By the π-system lemma, any two such measures must be equivalent. Hence the product measure $\mathbb{P}_1 \times \mathbb{P}_2$ is uniquely defined.

1.8 Further reading

This chapter gives a minimal introduction to measure-theoretic probability. Though this will be sufficient to get us through the course, this is no substitute for a good course on rigorous probability theory. I strongly encourage you to take the time to learn more about the foundations of this rich field.

There are many good books on probability theory; quite a few of them are very good indeed. Some (somewhat arbitrarily) selected textbooks where you can find many details on the topics of this chapter, and much more besides, are listed below.

It is hard to imagine a more lively introduction to probability theory than Williams' little blue book [Wil91]. A thorough and lucid development of probability theory can be found in the classic textbook by Billingsley [Bil86], while the textbook by Dudley [Dud02] puts a stronger emphasis on the analytic side of things. Finally, Kallenberg's monograph [Kal97] takes an original point of view on many topics in probability theory, but may be tough reading if you are learning the material for the first time.

CHAPTER 2

Conditioning, Martingales, and Stochastic Processes

The notion of conditional expectation is, in some sense, where probability theory gets interesting (and goes beyond pure measure theory). It allows us to introduce interesting classes of stochastic processes—Markov processes and martingales—which play a fundamental role in much of probability theory. Martingales in particular are ubiquitous throughout almost every topic in probability, even though this might be hard to imagine when you first encounter this topic.

We will take a slightly unusual route. Rather than introduce immediately the abstract definition of conditional expectations, we will start with the familiar discrete definition and build some of the key elements of the full theory in that context (particularly martingale convergence). This will be sufficient both to prove the Radon-Nikodym theorem, and to define the general notion of conditional expectation in a natural way. The usual abstract definition will follow from this approach, while you will get a nice demonstration of the power of martingale theory along the way.

2.1 Conditional expectations and martingales: a trial run

Discrete conditional expectations

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space (which we will fix until further notice), and consider two events $A, B \in \mathcal{F}$. You should be very comfortable with the notion of *conditional probability*: the probability that A occurs, given that B occurs, is $\mathbb{P}(A|B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$. Intuitively, if we repeat the experiment many times, but discard all of the runs where B did not occur, then $\mathbb{P}(A|B)$ is the fraction of the remaining runs in which A occured. Similarly, let X be a random variable. Then

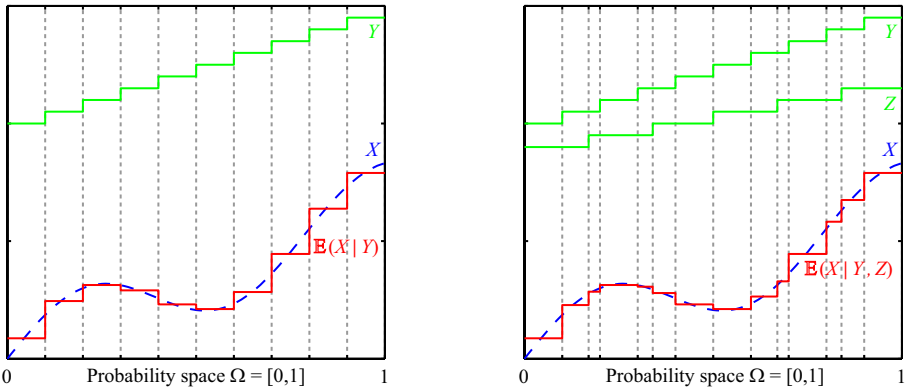

Figure 2.1. Illustration of the discrete conditional expectation on $([0,1], \mathcal{B}([0,1]), \mathbb{P})$, where $\mathbb{P}$ is the uniform measure. A random variable X is conditioned with respect to a discrete random variable Y (left) and with respect to two discrete random variables Y, Z (right). This amounts to averaging X (w.r.t. $\mathbb{P}$) over each bin in the partition generated by Y and Y, Z, respectively.

$\mathbb{E}(X|B) = \mathbb{E}(XI_B)/\mathbb{P}(B)$ is the expectation of X, given that B occurs. Intuitively, this is the mean value of our observations of X after we have discarded all runs of the experiment where B did not occur.

Often we are not so much interested in the conditional expectation with respect to an event, but rather with respect to some random variables which we have already observed. Suppose X, Y are two random variables where Y *takes a finite number of values* $y_1, \ldots, y_d$. In our experiment we observe Y, and we would like to determine the conditional mean of X as a function of our observation Y. For every possible outcome of Y we can separately determine the conditional mean $\mathbb{E}(X|Y = y_i)$, but it makes more sense in this context to think of the conditional mean as a true function of the observation: i.e., we should define the random variable $\mathbb{E}(X|Y) = f(Y)$, where $f(y_i) = \mathbb{E}(X|Y = y_i)$. $\mathbb{E}(X|Y)$ is called the *conditional expectation of* X *given* Y. You can think of this, if you wish, as a sort of estimator: $\mathbb{E}(X|Y)$ is a good estimate of X given Y, in some sense. We will make this idea more precise later on.

It is easy to extend this idea to a finite number of discrete random variables $Y^1, \ldots, Y^n$. Define the sets $A_{y^1,\ldots,y^n} = \{\omega \in \Omega : Y^1(\omega) = y^1, \ldots, Y^n(\omega) = y^n\}$. There are a finite number of these sets, as each of the Y^i only take a finite number of values; moreover, the sets are clearly disjoint and form a partition of Ω (in the sense that the union of all these sets is Ω). We now define $\mathbb{E}(X|Y^1, \ldots, Y^n)$ by setting $\mathbb{E}(X|Y^1, \ldots, Y^n)(\omega) = \mathbb{E}(X|A_{y^1,\ldots,y^n})$ for every $\omega \in A_{y^1,\ldots,y^n}$. This procedure is illustrated in figure 2.1. Note that once again $\mathbb{E}(X|Y^1, \ldots, Y^n) = f(Y^1, \ldots, Y^n)$, by construction, for some measurable function f.

We are now ready for a trivial but key insight. Our definition of the conditional expectation $\mathbb{E}(X|Y^1, \ldots, Y^n)$ does not actually depend on the values taken by the random variables $Y^1, \ldots, Y^n$; rather, it only depends on the partition $A_{y^1,\ldots,y^n}$ generated by these random variables. This partition encodes the maximal amount of

information that can be extracted by measuring these random variables: any event $B \in \sigma\{Y^1, \ldots, Y^n\}$ can be written as a union of the sets $A_{y^1,\ldots,y^n}$ (why?) Hence in reality we are not really conditioning on the random variables $Y^1, \ldots, Y^n$ themselves, but on the information contained in these random variables (which is intuitively precisely as it should be!) It should thus be sufficient, when we are calculating conditional expectations, to specify the σ-algebra generated by our observations rather than the observations themselves. This is what we will do from now on.

Definition 2.1.1. A σ-algebra $\mathcal{F}$ is said to be *finite* if it is generated by a finite number of sets $\mathcal{F} = \sigma\{A_1, \ldots, A_n\}$. Equivalently, $\mathcal{F}$ is finite if there is a finite partition of Ω such that every event in $\mathcal{F}$ is a union of sets in the partition (why are these equivalent?).

Definition 2.1.2. Let $X \in \mathcal{L}^1$ be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$, and let $\mathcal{G} \subset \mathcal{F}$ be a finite σ-algebra generated by the partition $\{A_k\}_{k=1,\ldots,n}$. Then

$$\mathbb{E}(X|\mathcal{G}) \equiv \sum_{k=1}^{n} \mathbb{E}(X|A_k)\, I_{A_k},$$

where $\mathbb{E}(X|A) = \mathbb{E}(XI_A)/\mathbb{P}(A)$ if $\mathbb{P}(A) > 0$ and we may define $\mathbb{E}(X|A)$ arbitrarily if $\mathbb{P}(A) = 0$. The conditional expectation thus defined is unique up to a.s. equivalence, i.e., any two random variables $Y, \tilde{Y}$ that satisfy the definition obey $Y = \tilde{Y}$ a.s. (why?).

Remark 2.1.3. $X \in \mathcal{L}^1$ ensures that $\mathbb{E}(X|A_k)$ is well defined and finite.

The conditional expectation defined here should be a completely intuitive concept. Unfortunately, extending it to σ-algebras which are not finite is not so straightforward. For example, suppose we would like to condition X on a random variable Y that is uniformly distributed on the unit interval. The quantity $\mathbb{E}(X|Y = y)$ is not well defined, however: $\mathbb{P}(Y = y) = 0$ for every $y \in [0, 1]$! In the finite case this was not a problem; if some sets in the partition had zero probability we just ignore them, and the resulting conditional expectation is still uniquely defined with probability one. In the continuous case, however, our definition *fails* with probability one (the problem being, of course, that there is an uncountable amount of trouble).

A look ahead

In laying the foundations of modern probability theory, one of the most important insights of Kolmogorov (the father of modern probability) was that the conditional expectation can be defined unambiguously even in the continuous case. Kolmogorov noticed that the discrete definition could be rephrased abstractly without mention of the finiteness of the σ-algebra, and that this abstract definition can serve as a meaningful definition of the conditional expectation in the continuous case. We could introduce this definition at this point and show that it reduces to the definition above for finite σ-algebras. To prove that the general definition is well posed, however, we need[1] the Radon-Nikodym theorem which we have not yet proved. It may also not

[1] This is not the only way to prove well posedness but, as we will see, there is a natural connection between the Radon-Nikodym theorem and the conditional expectation that makes this point of view worthwhile. We will comment on the other method of proving well posedness later on.

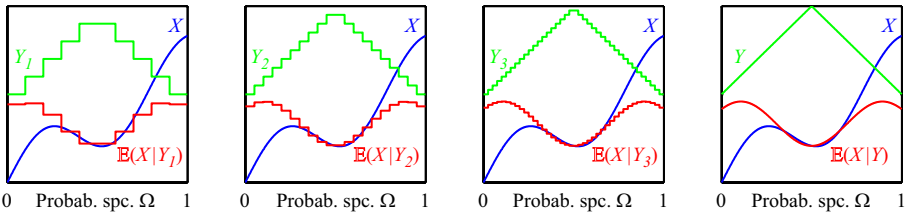

Figure 2.2. We would like to calculate $\mathbb{E}(X|Y)$, where $\mathbb{P}$ is the uniform measure on the interval, $X : [0,1] \to \mathbb{R}$ is some random variable and $Y : [0,1] \to \mathbb{R}$ is given by $Y(x) = x$ for $x \leq \frac{1}{2}$, $Y(x) = 1 - x$ for $x \geq \frac{1}{2}$. Intuitively we should have $\mathbb{E}(X|Y)(x) = \frac{1}{2}X(x) + \frac{1}{2}X(1-x)$, but definition 2.1.2 does not allow us to conclude this. However, a sequence of approximations covered by definition 2.1.2 appears to give rise to the expected result. But how to prove it?

be so appealing to just take for granted an abstracted definition, as it is not entirely obvious that it really encodes the desired concept.

A more natural idea, see figure 2.2, might be to try something like this. Suppose that we want to define $\mathbb{E}(X|Y)$, where Y is not necessarily finite-valued. However, we can easily find a sequence Y_n of finite-valued random variables such that $Y_n \to Y$ (e.g., using lemma 1.3.12): this is how we defined the expectation itself! One would thus think that we can define $\mathbb{E}(X|Y)$ as the limit of $\mathbb{E}(X|Y_n)$ as $n \to \infty$. To go down this path, we need to prove that this limit exists and is uniquely defined. Such problems are solved using martingale theory, which we can develop already in the simple framework of finite σ-algebras. This is what we will do in the remainder of this section. This route serves a dual purpose: we can provide an appealing definition of the conditional expectation, and on the way you can learn how martingales work in practice. Moreover, almost the same technique will allow us to prove the Radon-Nikodym theorem, thus tying all these topics together and showing how they relate.

At the end of the day, we will still introduce Kolmogorov's definition of the conditional expectation. This definition is much easier to use as it does not require taking limits, and makes many of the properties of the conditional expectation easy to prove. Hopefully, however, the intermediate story will help you get intuition for and practice in using conditional expectations and martingales.

Elementary properties of the conditional expectation

Some important properties of the conditional expectation are listed in the following lemma. These properties hold also for the general conditional expectation, but we will prove them here for the finite case (where many of the properties are trivial!)

Lemma 2.1.4. *Let $X, Y \in \mathcal{L}^1$ be random variables on $(\Omega, \mathcal{F}, \mathbb{P})$, let $\mathcal{G}, \mathcal{H} \subset \mathcal{F}$ be finite σ-algebras, and let $\alpha, \beta \in \mathbb{R}$.*

1. $\mathbb{E}(\alpha X + \beta Y|\mathcal{G}) = \alpha\,\mathbb{E}(X) + \beta\,\mathbb{E}(Y)$ *a.s.*
2. *If $X \geq 0$ a.s., then $\mathbb{E}(X|\mathcal{G}) \geq 0$ a.s.*

3. $\mathbb{E}(\mathbb{E}(X|\mathcal{G})) = \mathbb{E}(X)$.
4. *Tower property: if* $\mathcal{H} \subset \mathcal{G}$, *then* $\mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{H}) = \mathbb{E}(X|\mathcal{H})$ *a.s.*
5. *If* X *is* $\mathcal{G}$*-measurable, then* $\mathbb{E}(X|\mathcal{G}) = X$ *a.s.*
6. *If* X *is* $\mathcal{G}$*-measurable and* $XY \in \mathcal{L}^1$, *then* $\mathbb{E}(XY|\mathcal{G}) = X\,\mathbb{E}(Y|\mathcal{G})$ *a.s.*
7. *If* $\mathcal{H}$ *and* $\sigma\{X, \mathcal{G}\}$ *are independent, then* $\mathbb{E}(X|\sigma\{\mathcal{G}, \mathcal{H}\}) = \mathbb{E}(X|\mathcal{G})$ *a.s.*
8. *If* $\mathcal{H}$ *and* X *are independent, then* $\mathbb{E}(X|\mathcal{H}) = \mathbb{E}(X)$ *a.s.*
9. *Monotone and dominated convergence, Fatou's lemma, and Jensen's inequality all hold for conditional expectations also; e.g., if* $0 \le X_1 \le X_2 \le \cdots$ *a.s., then* $\mathbb{E}(X_n|\mathcal{G}) \nearrow \mathbb{E}(\lim_n X_n|\mathcal{G})$ *a.s. (monotone convergence).*

Proof.

1. Trivial.
2. Trivial.
3. $\mathbb{E}(\mathbb{E}(X|\mathcal{G})) = \sum_k \mathbb{E}(X|A_k)\mathbb{P}(A_k) = \sum_k \mathbb{E}(XI_{A_k}) = \mathbb{E}(XI_{\cup_k A_k}) = \mathbb{E}(X)$.
4. Let $\{A_k\}$ be the partition for $\mathcal{G}$ and $\{B_k\}$ be the partition for $\mathcal{H}$. Then
$$\mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{H}) = \sum_j \sum_k \mathbb{E}(X|A_k)\mathbb{P}(A_k|B_j)I_{B_j}.$$
But $\mathcal{H} \subset \mathcal{G}$ implies that every set B_j is the disjoint union of sets A_k, so $\mathbb{P}(A_k|B_j) = \mathbb{E}(I_{A_k}I_{B_j})/\mathbb{P}(B_j) = \mathbb{P}(A_k)/\mathbb{P}(B_j)$ if $A_k \subset B_j$, and zero otherwise. Hence
$$\mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{H}) = \sum_j \sum_{k:A_k\subset B_j} \frac{\mathbb{E}(XI_{A_k})}{\mathbb{P}(B_j)}\, I_{B_j} = \sum_j \frac{\mathbb{E}(XI_{B_j})}{\mathbb{P}(B_j)}\, I_{B_j} = \mathbb{E}(X|\mathcal{H}).$$
5. If X is $\mathcal{G}$-measurable, then $X^{-1}(B) \in \mathcal{G}$ for every $B \in \mathcal{B}(\mathbb{R})$. This implies that X must be constant on every set A_i in the partition for $\mathcal{G}$ (why?), so $X = \sum_i x_i I_{A_i}$ for some $x_i \in \mathbb{R}$. The result follows directly by plugging this into the definition.
6. As before, we may write $X = \sum_i x_i I_{A_i}$. Then
$$\mathbb{E}(XY|\mathcal{G}) = \sum_k \sum_i x_i\, \mathbb{E}(I_{A_i}Y|A_k)\, I_{A_k} = \sum_k x_k\, \mathbb{E}(Y|A_k)\, I_{A_k} = X\,\mathbb{E}(Y|\mathcal{G}),$$
where we have used (twice) that $I_{A_k}I_{A_i} = I_{A_k}$ if $k = i$, and zero otherwise.
7. Let $\{A_k\}$ be a partition for $\mathcal{G}$ and $\{B_k\}$ for $\mathcal{H}$. Then the sets $\{A_i \cap B_j\}$ are disjoint, and hence form a partition for $\sigma\{\mathcal{G}, \mathcal{H}\}$. We can thus write
$$\mathbb{E}(X|\mathcal{G}, \mathcal{H}) = \sum_{i,j} \frac{\mathbb{E}(XI_{A_i\cap B_j})}{\mathbb{P}(A_i \cap B_j)}\, I_{A_i\cap B_j} = \sum_{i,j} \frac{\mathbb{E}(XI_{A_i})\,\mathbb{P}(B_j)}{\mathbb{P}(A_i)\,\mathbb{P}(B_j)}\, I_{A_i}I_{B_j} = \mathbb{E}(X|\mathcal{G})$$
(with the convention $0/0 = 0$), where we have used the independence of XI_{A_i} and I_{B_j}.
8. Use the previous result with $\mathcal{G} = \{\varnothing, \Omega\}$.
9. Trivial. □

Discrete time stochastic processes and filtrations

A *stochastic process* is just a collection of random variables $\{X_t\}$, indexed by time t. In later chapters we will most often work in continuous time $t \in [0, \infty[$, but for the time being we will concentrate on the discrete time case, i.e., $t = 0, 1, 2, \ldots$ In this form, a stochastic process is nothing to get excited about. After all, a discrete time stochastic process is just a sequence of random variables $\{X_n\}$—we have encountered plenty such sequences already. What do we gain by interpreting the index n as "time"?

Stochastic processes start leading a life of their own once we build a notion of time into our probability space. In our elementary space $(\Omega, \mathcal{F}, \mathbb{P})$, the σ-algebra $\mathcal{F}$ is the set of all yes-no questions that could be asked (and answered) during the course of an experiment. However, not all such questions can be answered by some fixed time. For example, suppose we flip coins, where X_n is the outcome of the nth coin flip. Then at time n, we know the answer to the question *did the flips up to time n come up heads more often than tails?*, but not to the question *will we flip more heads than tails before time $N > n$?* (we could calculate the probability of the latter event, but we can never *know* its outcome at time n!) To build the notion of time into our probability space, we need to specify which sub-σ-algebra of questions in $\mathcal{F}$ can be answered by time n. If we label this σ-algebra by $\mathcal{F}_n$, we obtain the following notion.

Definition 2.1.5. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A (discrete time) *filtration* is an increasing sequence $\{\mathcal{F}_n\}$ of σ-algebras $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}$. The quadruple $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}, \mathbb{P})$ is called a *filtered probability space*.

Note that the sequence $\mathcal{F}_n$ must be increasing—a question that can be answered by time n can also be answered at any later time. We can now introduce a notion of causality for stochastic processes.

Definition 2.1.6. Let $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}, \mathbb{P})$ be a filtered probability space. A stochastic process $\{X_n\}$ is called ($\mathcal{F}_n$-) *adapted* if X_n is $\mathcal{F}_n$-measurable for every n, and is called ($\mathcal{F}_n$-) *predictable* if X_n is $\mathcal{F}_{n-1}$-measurable for every n.

Hence if $\{X_n\}$ is adapted, then X_n represents a measurement of something in the past or present (up to and including time n), while in the predictable case X_n represents a measurement of something in the past (before time n). Note how the notion of time is now deeply embedded in our probabilistic model—time is much more than just an index in a sequence of random variables!

Conversely, if we have some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a sequence of random variables $\{X_n\}$, we can use this sequence to generate a filtration:

Definition 2.1.7. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\{X_n\}$ be a stochastic process. The *filtration generated by* $\{X_n\}$ is defined as $\mathcal{F}_n^X = \sigma\{X_0, \ldots, X_n\}$, and the process $\{X_n\}$ is $\mathcal{F}_n^X$-adapted by construction.

In practice, filtrations are very often generated in this way. Once we have such a filtration, we can use it as before to define a notion of time. It turns out that this is a very fruitful point of view. Even if we are just dealing with a sequence $\{X_n\}$, where n has no relation to the physical notion of time (e.g., the sequence of approximations $\mathbb{E}(X|Y_n)$ that we wish to use to define $\mathbb{E}(X|Y)$), it will pay off to think of

this sequence as an adapted stochastic process. On the other hand, many stochastic processes used to model random signals or natural phenomena do have an associated physical notion of time, which is faithfully encoded using the concept of a filtration.

Martingales

A martingale is a very special type of stochastic process.

Definition 2.1.8. A stochastic process $\{X_n\}$ is said to be an $\mathcal{F}_n$*-martingale* if it is $\mathcal{F}_n$-adapted and satisfies $\mathbb{E}(X_n|\mathcal{F}_m) = X_m$ a.s. for every $m \leq n$. (If the filtration is obvious, e.g., on a filtered probability space, we will just say that X_n is a *martingale*).

Remark 2.1.9. We have not yet defined the conditional expectation for anything but finite σ-algebras. Thus until further notice, *we assume that $\mathcal{F}_n$ is finite for every n.* In particular, this means that if X_n is $\mathcal{F}_n$-adapted, then every X_n is a finite-valued random variable. This will be sufficient machinery to develop the general theory.

How should you interpret a martingale? The basic idea (and the pretty name) comes from gambling theory. Suppose we play a sequence of games at a casino, in each of which we can win or lose a certain amount of money. Let us denote by X_n our total winnings after the nth game: i.e., X_0 is our starting capital, X_1 is our starting capital plus our winnings in the first game, etc. We do not assume that the games are independent. For example, we could construct some crazy scheme where we play poker in the nth game if we have won an even number of times in the past, and we play blackjack if we have won an odd number of times in the past. As poker and blackjack give us differently distributed winnings, our winnings $X_n - X_{n-1}$ in the nth game will then depend on all of the past winnings $X_0, \ldots, X_{n-1}$.

If the game is *fair*, however, then we should make no money on average in any of the games, regardless of what the rules are. After all, if we make money on average then the game is unfair to the casino, but if we lose money on average the game is unfair towards us (most casinos operate in the latter mode). As such, suppose we have made X_m dollars by time m. If the game is fair, then our *expected* winnings at any time in the future, given the history of the games to date, should equal our current capital: i.e., $\mathbb{E}(X_n|\sigma\{X_0, \ldots, X_m\}) = X_m$ for any $n \geq m$. This is precisely the definition of an ($\mathcal{F}_n^X$-)martingale. Hence we can interpret a martingale as the winnings in a sequence of fair games (which may have arbitrarily complicated rules).

You might be surprised that such a concept has many far-reaching consequences. Indeed, martingale techniques extend far beyond gambling, and pervade almost all aspects of modern probability theory. It was the incredible insight of J. L. Doob [Doo53] that martingales play such a fundamental role. There are many reasons for this. First, martingales have many special properties, some of which we will discuss in this chapter. Second, martingales show up naturally in many situations which initially appear to have little to do with martingale theory. The following simple result (which we will not need during the rest of this course) gives a hint as to why this could be the case.

Lemma 2.1.10 (Doob decomposition). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\mathcal{F}_n$ be a (finite) filtration and let $\{X_n\}$ be $\mathcal{F}_n$-adapted with $X_n \in \mathcal{L}^1$ for every n. Then*

$X_n = X_0 + A_n + M_n$ a.s., where $\{A_n\}$ is $\mathcal{F}_n$-predictable and $\{M_n\}$ is an $\mathcal{F}_n$-martingale with $M_0 = 0$. Moreover, this decomposition is unique.

Proof. Let $A_n = \sum_{k=1}^{n} \mathbb{E}(X_k - X_{k-1}|\mathcal{F}_{k-1})$; A_n is well defined ($X_k \in \mathcal{L}^1$) and predictable. We claim that $M_n = X_n - X_0 - A_n$ is a martingale (clearly $M_0 = 0$). But this follows from

$$M_n = X_n - X_0 - A_n = \sum_{k=1}^{n} \{X_k - X_{k-1} - \mathbb{E}(X_k - X_{k-1}|\mathcal{F}_{k-1})\}$$

using lemma 2.1.4 (why?) To prove uniqueness, suppose that $\tilde{M}_n$ and $\tilde{A}_n$ were another martingale (with $\tilde{M}_0 = 0$) and predictable process, respectively, such that $X_n = X_0 + \tilde{A}_n + \tilde{M}_n$ a.s. Then evidently $\tilde{A}_n - A_n = M_n - \tilde{M}_n$ a.s. But the left hand side is $\mathcal{F}_{n-1}$-measurable, so using the martingale property $\tilde{A}_n - A_n = \mathbb{E}(\tilde{A}_n - A_n|\mathcal{F}_{n-1}) = \mathbb{E}(M_n - \tilde{M}_n|\mathcal{F}_{n-1}) = 0$ a.s. Hence $A_n = \tilde{A}_n$ a.s., and consequently $M_n = \tilde{M}_n$ a.s. as well. □

Remark 2.1.11. Note that *any* discrete time stochastic process can be decomposed into a martingale part and a predictable part (provided that $X_n \in \mathcal{L}^1$ for all n). This result still holds when the $\mathcal{F}_n$ are not finite, but is not true in continuous time. Nonetheless almost all processes of interest have such a decomposition. For example, we will see in later chapters that the solution of a stochastic differential equation can be written as the sum of a martingale and a process which is differentiable in time. For similar reasons, martingales play an important role in the general theory of Markov processes. As this is an introductory course, we will not attempt to lay down these theories in such generality; the purpose of this interlude was to give you an idea of how martingales can emerge in seemingly unrelated problems.

Many results about martingales are proved using the following device.

Definition 2.1.12. Let $\{M_n\}$ be a martingale and $\{A_n\}$ be a predictable process. Then $(A \cdot M)_n = \sum_{k=1}^{n} A_k(M_k - M_{k-1})$, the *martingale transform* of M by A, is again a martingale, provided that A_n and $(A \cdot M)_n$ are in $\mathcal{L}^1$ for all n (why?).

Let us once again give a gambling interpretation. We play a sequence of games; before every game, we may stake a certain amount of money. We now interpret the martingale M_n not as our total winnings at time n, but as our total winnings if we were to stake one dollar on each game. For example, if we stake A_1 dollars on the first game, then we actually win $A_1(M_1 - M_0)$ dollars. Consequently, if we stake A_k dollars on the kth game, then our total winnings after the nth game are given by $X_n = X_0 + (A \cdot M)_n$. Note that it is important for A_n to be predictable: we have to place our bet *before* the game is played, so our decision on how much money to stake can only depend on the past (obviously we could always make money, if we knew the outcome of the game in advance!) Other than that, we are free to choose an arbitrarily complicated gambling strategy A_n (our decision on how much to stake on the nth game can depend arbitrarily on what happened in previous games). The fact that X_n is again a martingale says something we know intuitively—there is no "reasonable" gambling strategy that allows us to make money, on average, on a fair game.

We are now ready to prove of of the key results on martingales—the *martingale convergence theorem*. With that bit of machinery in hand, we will be able to prove the Radon-Nikodym theorem and to extend our definition of the conditional expectation.

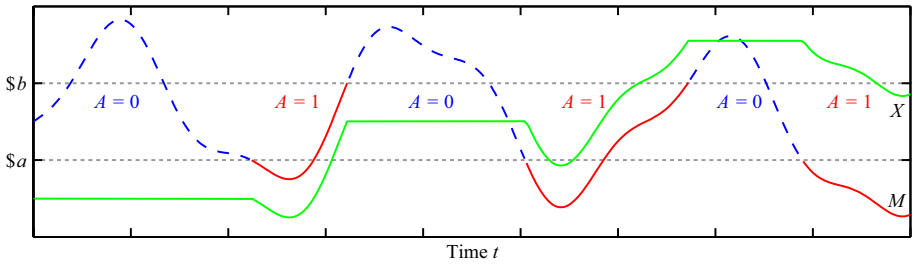

Figure 2.3. We stake bets on a martingale M as follows: we start betting (with a fixed stake $A = 1$) when M is first below a, we stop betting ($A = 0$) when M next exceeds b, and repeat. The periods when we are betting are shown in red and when we are not in blue. Our total winnings X are in green. At the final time T, our winnings can evidently not exceed $b - a$ times the number of upcrossings (two) minus $(a - M_T)^+$. (Figure adapted from [Wil91].)

The martingale convergence theorem

Let us go to the casino. We join the roulette table, where a lone soul is betting his money away. Our new friend has a simple strategy: he stakes one dollar on every game. Having selected the only fair casino in the world, our friend makes no money on average. We, on the other hand, think we can do better by playing the following strategy. We do not stake any money until our friend's capital sinks below a certain amount of dollars a. We then stake one dollar per game until our friend's capital exceeds $b > a$ dollars. At this point, we stop staking money until our friend's capital has dropped below a dollars again, and repeat. As long as our friend's luck keeps changing, we will repeatedly make $b - a$ dollars and thus get very rich; or not?

Previously, we concluded that the winnings obtained through any gambling strategy is again a martingale (the $\mathcal{L}^1$ condition is clearly satisfied here, as our gambling strategy A_n is bounded). Hence at any time, we should have made no money on average even with our smart alec strategy. But how can this be? Suppose that at time n, our friend has had k reversals of fortune, so we have made $k(b-a)$ dollars by starting to stake low and stopping to stake high. Somehow, this profit must be offset by a loss so that our total winnings will average to zero. The only winnings that we have not taken into account, however, are procured if our friend's wealth has actually hit its low point a for the $(k+1)$th time, but has not yet hit the high point b for the $(k+1)$th time. Once we hit the level b again (in the future) we will have made another $b - a$ dollars, but we could actually make a significant loss before this time (see figure 2.3). The only way that our winnings at time n can average to zero, is if the expected loss incurred since the last time we started staking money in the game equals $k(a - b)$.

Now suppose that our friend keeps having more and more reversals of fortune; that is, we repeatedly make $b - a$ dollars. The only way that this can happen is if our intermediate losses get larger and larger (after all, everything must still average to zero). If, on the other hand, we know that our friend's winnings are bounded in some sense—for example, the casino could refuse to let him play, if he is too much

in debt—then the latter cannot happen. Hence there is only one logical conclusion in this case: evidently our friend's winnings can cross a and b only a *finite* number of times; otherwise we could make money on average by playing a predictable gambling strategy in a fair game. But this must be true for every value of a and b (these were only used in our gambling strategy, they do not determine our friend's winnings!), so we come to a very interesting conclusion: if a martingale M_n is bounded, then it cannot fluctuate forever; in other words, *it must converge* to some random variable M_∞. We have basically proved the martingale convergence theorem.

Let us now make these ideas precise.

Lemma 2.1.13 (Doob's upcrossing lemma). *Let $\{M_n\}$ be a martingale, and denote by $U_n(a,b)$ the number of upcrossings of $a < b$ up to time n: that is, $U_n(a,b)$ is the number of times that M_k crosses from below a to above b before time n. Then we have $\mathbb{E}(U_n(a,b)) \le \mathbb{E}((a-M_n)^+)/(b-a)$.*

Proof. Define the following gambling strategy (see figure 2.3). Let $A_0 = 0$, and set $A_k = 1$ if either $A_{k-1} = 1$ and $M_{k-1} < b$, or if $A_{k-1} = 0$ and $M_{k-1} \le a$, and set $A_k = 0$ otherwise. Clearly A_k is bounded and predictable, so $X_n = (A \cdot M)_n$ is a martingale (this is the winnings process of figure 2.3). We can evidently estimate $X_n \ge (b-a)\,U_n(a,b) - (a-M_n)^+$; the first term is the number of upcrossings times the winnings per upcrossing, while the second term is a lower bound on the loss incurred after the last upcrossing before time n. As X_n is a martingale, however, $\mathbb{E}(X_n) = X_0 = 0$, and the result follows directly. □

Theorem 2.1.14 (Martingale convergence). *Let $\{M_n\}$ be an $\mathcal{F}_n$-martingale such that one of the following hold: (a) $\sup_n \mathbb{E}(|M_n|) < \infty$; or (b) $\sup_n \mathbb{E}((M_n)^+) < \infty$; or (c) $\sup_n \mathbb{E}((M_n)^-) < \infty$. Then there exists an $\mathcal{F}_\infty$-measurable random variable $M_\infty \in \mathcal{L}^1$, where $\mathcal{F}_\infty = \sigma\{\mathcal{F}_n : n = 1,2,\ldots\}$, such that $M_n \to M_\infty$ a.s.*

Proof. We can assume without loss of generality that $M_0 = 0$ (otherwise just consider the martingale $N_n = M_n - M_0$). Let us first show that the three conditions (a)–(c) are equivalent. Note that $0 = \mathbb{E}(M_n) = \mathbb{E}((M_n)^+) - \mathbb{E}((M_n)^-)$, so $\mathbb{E}((M_n)^+) = \mathbb{E}((M_n)^-)$. Moreover $\mathbb{E}(|M_n|) = \mathbb{E}((M_n)^+) + \mathbb{E}((M_n)^-)$, so clearly (a)–(c) are equivalent.

Next, note that $(a-M_n)^+ \le |a-M_n| \le |a| + |M_n|$. Hence for any n, we can bound $\mathbb{E}(U_n(a,b)) \le (|a| + \sup_n \mathbb{E}(|M_n|))/(b-a) < \infty$. But then letting $n \to \infty$ and using monotone convergence, we find that $\mathbb{E}(U_\infty(a,b)) < \infty$. Thus apparently, M_n can only ever have a finite number of upcrossings of $a < b$ for fixed a and b.

Now there are three possibilities. Either $M_n(\omega)$ converges to some finite value as $n \to \infty$; or $M_n(\omega)$ converges to $+\infty$ or $-\infty$; or $M_n(\omega)$ does not have a limit (its limit superior and inferior are different). In the latter case, there must be some *rational* numbers $a < b$ that are crossed by $M_n(\omega)$ infinitely many times (choose $\liminf_n M_n(\omega) < a < b < \limsup M_n(\omega)$). But

$$\begin{aligned}\mathbb{P}(\exists\, a,b \in \mathbf{Q},\ a<b \text{ s.t. } M_n \text{ crosses } a,b \text{ infinitely often}) \\ \le \sum_{a,b\in\mathbf{Q}} \mathbb{P}(M_n \text{ crosses } a,b \text{ infinitely often}) = \sum_{a,b\in\mathbf{Q}} \mathbb{P}(U_\infty(a,b)=\infty) = 0.\end{aligned}$$

Hence we have established that M_n converges a.s. to some M_∞. But by Fatou's lemma $\mathbb{E}(|M_\infty|) \le \liminf_n \mathbb{E}(|M_n|) < \infty$, so M_∞ must be a.s. finite and even in $\mathcal{L}^1$. □

Remark 2.1.15. Note that we are still in the setting where $\mathcal{F}_n$ is finite for all n. However, nothing in the proofs of these results uses (or is hindered by) this fact, and the proofs will carry over immediately to the general case.

Towards a general conditional expectation

Let X, Y be two random variables. We would like to give meaning to $\mathbb{E}(X|Y)$, but Y does not take a finite number of values (ultimately we wish to define $\mathbb{E}(X|\mathcal{G})$ for general $\mathcal{G}$, but we will go into this shortly). However, we can find a sequence of $\sigma\{Y\}$-measurable discrete approximations Y_n that converge to Y, as in lemma 1.3.12, and for each n we can define $M_n = \mathbb{E}(X|Y_1, \ldots, Y_n)$ using our existing theory. We now claim that M_n converges a.s., so we may define $\mathbb{E}(X|Y) \equiv \lim_n M_n$.

Why is this true? The key point is that the sequence M_n *is a martingale* with respect to the natural filtration $\mathcal{F}_n = \sigma\{Y_1, \ldots, Y_n\}$. To see this, note that M_n is clearly adapted. Moreover $\mathbb{E}(M_n|\mathcal{F}_m) = \mathbb{E}(\mathbb{E}(X|\mathcal{F}_n)|\mathcal{F}_m) = \mathbb{E}(X|\mathcal{F}_m) = M_m$ for any $m \leq n$ by the tower property of the conditional expectation, so M_n is a martingale. Finally, by Jensen's inequality $|M_n| \leq \mathbb{E}(|X| \ |\mathcal{F}_n)$, so $\mathbb{E}(|M_n|) \leq \mathbb{E}(|X|)$ for all n. Hence the boundedness condition for the martingale convergence theorem is satisfied, and we conclude that $M_n \to M_\infty \equiv \mathbb{E}(X|Y)$ a.s. (see figure 2.2).

Of course, we are not done yet: we still need to convince ourselves that our definition of $\mathbb{E}(X|Y)$ does not depend on the choice of approximating sequence Y_n. But first, we will take a little detour into the proof of the Radon-Nikodym theorem.

2.2 The Radon-Nikodym theorem revisited

Separable σ-algebras

Let us briefly recap where we left off. We have a notion of conditional expectations that works for finite σ-algebras; in this framework, we can define discrete martingales. We would like to use martingale convergence to extend the discrete theory to the continuous case. We thus need something like the following concept.

Definition 2.2.1. A σ-algebra $\mathcal{F}$ is called *separable* if there exists a filtration $\{\mathcal{F}_n\}$ of discrete σ-algebras $\mathcal{F}_n \subset \mathcal{F}$ such that $\mathcal{F} = \sigma\{\mathcal{F}_n : n = 1, 2, \ldots\}$. Equivalently (why?), $\mathcal{F}$ is separable if $\mathcal{F} = \sigma\{A_n\}$ for a countable collection of events $A_n \in \mathcal{F}$.

Remark 2.2.2. In this course, we will never go beyond separable σ-algebras; we will be content to prove, e.g., the Radon-Nikodym theorem, for separable σ-algebras only. In fact, many results (such as the Radon-Nikodym theorem) can be extended to the non-separable case by approximating a non-separable σ-algebra by separable σ-algebras; see [Wil91, p. 147–149] for such an argument. This does not add any intuition, however, so we will not bother to do this.

A large number of the σ-algebras encountered in practice—all the σ-algebras which you will encounter in these notes fall in this category!—turn out to be separable. There are certain cases where non-separable σ-algebras become important (if you know about such things, think of the tail σ-algebra of a sequence of i.i.d. random variables), but we will not run into them in this course.

Example 2.2.3. Let Y be a random variable; then $\sigma\{Y\}$ is separable: set $\mathcal{F} = \sigma\{\mathcal{F}_n : n = 1, 2, \ldots\}$, where $\mathcal{F}_n = \sigma\{Y^1, \ldots, Y^n\}$ with Y^k as in lemma 1.3.12.

Example 2.2.4. If X_n is a sequence of random variables, then $\mathcal{F} = \sigma\{X_n\}$ is separable: approximating every X_n by X_n^k, choose $\mathcal{F}_n = \sigma\{X_m^k : m, k = 1, \ldots, n\}$.

Example 2.2.5. Let $\{X_t\}_{t\in[0,\infty[}$ be a continuous time stochastic process such that $t \mapsto X_t(\omega)$ is continuous for every ω. Then $\mathcal{F} = \sigma\{X_t : t \in [0, \infty[\}$ is separable. To see this, note that by continuity $t \mapsto X_t$ is completely known if we know it for a dense set of times (e.g., the dyadic rationals). Hence approximating X_t by a sequence X_t^k for every t, we can use $\mathcal{F}_n = \sigma\{X_t^k : k = 1, \ldots, n,\ t = \ell 2^{-n},\ \ell = 0, \ldots, n2^n\}$.

Proof of the Radon-Nikodym theorem

We are finally going to prove the Radon-Nikodym theorem, albeit for separable σ-algebras. You can easily guess what we are going to do: we will prove the theorem for finite σ-algebras (trivial), then take a limit.

Lemma 2.2.6 (Finite Radon-Nikodym). *Let $\mathcal{F}$ be a finite σ-algebra and let $\mathbb{Q} \ll \mathbb{P}$ be probability measures on $\mathcal{F}$. Then there is an a.s. unique $\mathcal{F}$-measurable random variable $f = d\mathbb{Q}/d\mathbb{P}$ such that $\mathbb{Q}(A) = \mathbb{E}_{\mathbb{P}}(I_A f)$ for every $A \in \mathcal{F}$.*

Proof. Let $\{A_n\}$ be a partition of Ω that generates $\mathcal{F}$. Define $f(\omega) = \mathbb{Q}(A_k)/\mathbb{P}(A_k)$ for all $\omega \in A_k$, where we may assign an arbitrary value if $\mathbb{P}(A_k) = 0$. Clearly f is $\mathcal{F}$-measurable, $\mathbb{Q}(A_k) = \mathbb{E}_{\mathbb{P}}(I_{A_k} f)$ when $\mathbb{P}(A_k) > 0$, whereas both sides are zero when $\mathbb{P}(A_k) = 0$ (note that $\mathbb{Q} \ll \mathbb{P}$ is crucial for this to hold!) As any set $A \in \mathcal{F}$ can be written as the union of A_ks, we find that $\mathbb{Q}(A) = \mathbb{E}_{\mathbb{P}}(I_A f)$ for any $A \in \mathcal{F}$. This settles existence of $d\mathbb{Q}/d\mathbb{P}$.

Uniqueness is essentially trivial. Let $\tilde{f}$ be another $\mathcal{F}$-measurable function; then $\tilde{f}$ must be constant on all A_k. If $\tilde{f} \neq f$ on a set A_k with $\mathbb{P}(A_k) > 0$, then $\mathbb{E}_{\mathbb{P}}(I_{A_k}\tilde{f}) \neq \mathbb{Q}(A)$. So we may only change f on a set A_k of measure zero; but then $f = \tilde{f}$ a.s. □

A reminder:

Theorem 1.6.12 (Radon-Nikodym). *Suppose $\mathbb{Q} \ll \mathbb{P}$ are two probability measures on the space $(\Omega, \mathcal{F})$. Then there exists a nonnegative $\mathcal{F}$-measurable function f with $\mathbb{E}_{\mathbb{P}}(f) = 1$, such that $\mathbb{Q}(A) = \mathbb{E}_{\mathbb{P}}(I_A f)$ for every $A \in \mathcal{F}$, and f is unique in the sense that if f' is another $\mathcal{F}$-measurable function with this property, then $f' = f$ $\mathbb{P}$-a.s.*

Assume $\mathcal{F} = \sigma\{\mathcal{F}_n : n = 1, 2, \ldots\}$. Applying lemma 2.2.6 to the $\mathcal{F}_n$, we obtain a sequence f_n of finite Radon-Nikodym derivatives. We now proceed in three steps.

1. We will find a candidate Radon-Nikodym derivative for $\mathcal{F}$ by taking the limit $f = \lim_n f_n$. To this end we show that $\{f_n\}$ is an $\mathcal{L}^1$-bounded martingale, so that convergence is guaranteed by the martingale convergence theorem.

2. We must show that f thus defined indeed satisfies $\mathbb{Q}(A) = \mathbb{E}_{\mathbb{P}}(I_A f)$ for every $A \in \mathcal{F}$. This requires some creative use of the limit theorems for random variables, and the Dynkin π-system lemma 1.7.3.

3. We must show uniqueness. It then follows that the limit $\lim_n f_n$ is independent of the choice of discretization $\{\mathcal{F}_n\}$ (which is not obvious from the outset!)

Let us get to work.

Proof. Throughout the proof, $\mathbb{E}$ denotes the expectation with respect to $\mathbb{P}$.

1. Let f_n be the Radon-Nikodym derivative obtained by applying lemma 2.2.6 to $\mathcal{F}_n$. Then f_n is a discrete random variable (as $\mathcal{F}_n$ is finite). We claim that the sequence $\{f_n\}$ is an $\mathcal{F}_n$-martingale. To see this, let $\{A_k\}$ be a partition that generates $\mathcal{F}_n$ and $\{B_j\}$ be a partition that generates $\mathcal{F}_m$, $m < n$. Then every B_j is a union of sets A_k. Thus

$$\mathbb{E}(f_n|B_j) = \frac{1}{\mathbb{P}(B_j)} \sum_{k:A_k \subset B_j} \mathbb{E}(f_n I_{A_k}) = \frac{1}{\mathbb{P}(B_j)} \sum_{k:A_k \subset B_j} \mathbb{Q}(A_k) = \frac{\mathbb{Q}(B_j)}{\mathbb{P}(B_j)}.$$

   Hence evidently $\mathbb{E}(f_n|\mathcal{F}_m) = \sum_j \mathbb{E}(f_n|B_j) I_{B_j} = f_m$. But note that f_n is clearly nonnegative for all n, so the boundedness condition of the martingale convergence theorem holds trivially. Hence f_n converges $\mathbb{P}$-a.s., and we can define $f = \lim_n f_n$. But as $\mathbb{Q} \ll \mathbb{P}$, we find that $f_n \to f$ $\mathbb{Q}$-a.s. as well. This will be crucial below.

2. The hardest part here is to show that $\mathbb{E}(f) = 1$. Let us complete the argument assuming that this is the case. Note that $\mathcal{G} = \bigcup_n \mathcal{F}_n$ is a π-system. We would like to show that $\mathbb{Q}(A) = \mathbb{P}(I_A f)$ for all $A \in \mathcal{F}$; but as $\mathbb{E}(f) = 1$, both sides are valid probability measures and it suffices to check this for $A \in \mathcal{G}$ (by the π-system lemma 1.7.3). Now for any $A \in \mathcal{G}$, there exists by definition an m such that $A \in \mathcal{F}_m$. Hence $\mathbb{P}(I_A f_n) = \mathbb{Q}(A)$ for $n \geq m$ by lemma 2.2.6. Using Fatou's lemma,

$$\mathbb{P}(I_A f) = \mathbb{P}(\liminf_n I_A f_n) \leq \liminf_n \mathbb{P}(I_A f_n) = \mathbb{Q}(A) \qquad \forall A \in \mathcal{G}.$$

   But we obtain the inequality in the reverse direction by applying this expression to A^c and using $\mathbb{E}(f) = 1$. Hence indeed $f = d\mathbb{Q}/d\mathbb{P}$, provided we can show that $\mathbb{E}(f) = 1$.

   To show $\mathbb{E}(f) = 1$, we would usually employ the dominated convergence theorem (as $\mathbb{E}(f_n) = 1$ for all n). Unfortunately, it is not obvious how to dominate $\{f_n\}$. To circumvent this, we rely on another useful trick: a truncation argument. Define

$$\varphi_n(x) = \begin{cases} 1 & x \leq n, \\ n+1-x & n \leq x \leq n+1, \\ 0 & x \geq n+1. \end{cases}$$

   The function φ_n is continuous for every n and $\varphi_n \nearrow 1$. Moreover, $f_m \varphi_n(f_m)$ is bounded by $n+1$ for all m, n. But $\mathbb{E}(f_m \varphi_n(f_m)) = \mathbb{E}_{\mathbb{Q}}(\varphi_n(f_m))$ by lemma 2.2.6, as $\varphi_n(f_m)$ is $\mathcal{F}_m$-measurable. Letting $m \to \infty$ using dominated convergence (which we can apply now, having truncated the integrands to be bounded!), we find $\mathbb{E}(f\varphi_n(f)) = \mathbb{E}_{\mathbb{Q}}(\varphi_n(f))$. Note that it is crucial here that $f_n \to f$ both $\mathbb{P}$-a.s. and $\mathbb{Q}$-a.s.! Finally, let $n \to \infty$ using monotone convergence; this gives $\mathbb{E}(f) = 1$.

3. Suppose that f and $\tilde{f}$ are both $\mathcal{F}$-measurable and satisfy $\mathbb{E}(I_A f) = \mathbb{E}(I_A \tilde{f}) = \mathbb{Q}(A)$ for all $A \in \mathcal{F}$. Define $A_+ = \{\omega : f(\omega) > \tilde{f}(\omega)\}$ and $A_- = \{\omega : \tilde{f}(\omega) > f(\omega)\}$. Then $A_+, A_- \in \mathcal{F}$. But if $\mathbb{P}(A_+) > 0$, then we would have $\mathbb{E}(I_{A_+}(f - \tilde{f})) > 0$ which is ruled out by assumption, so $\mathbb{P}(A_+) = 0$. Similarly $\mathbb{P}(A_-) = 0$, so $f = \tilde{f}$ a.s. ∎

Existence and uniqueness of the conditional expectation

Let us return to the seemingly unrelated issue of defining the conditional expectation.

Definition 2.2.7. Let $\mathcal{F}$ be a separable σ-algebra, i.e., $\mathcal{F} = \sigma\{\mathcal{F}_n : n = 1, 2, \ldots\}$ with $\mathcal{F}_n$ finite. Let $X \in \mathcal{L}^1$. Then we define $\mathbb{E}(X|\mathcal{F}) \equiv \lim_n \mathbb{E}(X|\mathcal{F}_n)$.

You can easily convince yourself that the sequence $M_n = \mathbb{E}(X|\mathcal{F}_n)$ is a (discrete) $\mathcal{F}_n$-martingale, and moreover $\sup_n \mathbb{E}(|M_n|) \le \mathbb{E}(|X|)$ as before. Hence the limit as $n \to \infty$ does indeed exist, and is even in $\mathcal{L}^1$, by the martingale convergence theorem. However, we are unsatisfied, because we might well get a different answer if we used a different discretization sequence $\{\mathcal{F}_n : n = 1, 2, \ldots\}$.

Consider, however, the following idea. Choose for simplicity (we will see shortly why!) an X such that $X \ge 0$ a.s. and $\mathbb{E}(X) = 1$. Then for every $\mathcal{F}_n$, we can write

$$\mathbb{E}(X|\mathcal{F}_n) = \sum_k \mathbb{E}(X|A_k) I_{A_k} = \sum_k \frac{\mathbb{E}(I_{A_k} X)}{\mathbb{P}(A_k)} I_{A_k},$$

where A_k is a partition that generated $\mathcal{F}_n$. But this is just a Radon-Nikodym derivative in disguise: if we define the probability measure $\mathbb{Q}(A) = \mathbb{P}(I_A X)$, then

$$\mathbb{E}(X|\mathcal{F}_n) = \sum_k \frac{\mathbb{Q}(A_k)}{\mathbb{P}(A_k)} I_{A_k} = \frac{d\mathbb{Q}|_{\mathcal{F}_n}}{d\mathbb{P}|_{\mathcal{F}_n}},$$

where we write $\mathbb{Q}|_{\mathcal{F}_n}$ to signify that we have restricted the measure $\mathbb{Q}$ to $\mathcal{F}_n$ (i.e., apply lemma 2.2.6 with $\mathcal{F} = \mathcal{F}_n$), and similarly for $\mathbb{P}|_{\mathcal{F}_n}$. In particular, if we let $n \to \infty$, then our proof of the Radon-Nikodym theorem shows that

$$\mathbb{E}(X|\mathcal{F}) = \lim_n \frac{d\mathbb{Q}|_{\mathcal{F}_n}}{d\mathbb{P}|_{\mathcal{F}_n}} = \frac{d\mathbb{Q}|_{\mathcal{F}}}{d\mathbb{P}|_{\mathcal{F}}}.$$

Thus apparently, there is a fundamental connection between the notion of a Radon-Nikodym derivative and the notion of a conditional expectation! This immediately resolves our uniqueness problem: as we have seen (this was not difficult at all) that the Radon-Nikodym derivative does not depend on the discretization $\{\mathcal{F}_n\}$, it is now clear that neither does our definition the conditional expectation. We have thus finally come to the conclusion that definition 2.2.7, which we have been hinting at (not so subtly) for almost the entire chapter to date, really does make sense.

Remark 2.2.8. The choices $X \ge 0$ a.s. and $\mathbb{E}(X) = 1$ are in no way a restriction; these merely make $\mathbb{Q}$ a probability measure, which was however not essential to the argument. We can always rescale $X \in \mathcal{L}^1$ so that it has unit expectation, while we can always express any X as $X^+ - X^-$. As all the discrete conditional expectations are linear and all the limits exist by martingale convergence, it is immediate that $\mathbb{E}(X|\mathcal{F})$ is also linear with respect to X; hence everything extends directly to arbitrary $X \in \mathcal{L}^1$.

2.3 Conditional expectations and martingales for real

The Kolmogorov definition

As we saw in the last section, the conditional expectation $\mathbb{E}(X|\mathcal{F})$ can be defined as the Radon-Nikodym derivative of the measure $\mathbb{Q}(A) = \mathbb{E}(I_A X)$ with respect to the measure $\mathbb{P}$ (at least for $X \geq 0$ such that $\mathbb{E}(X) = 1$; then extend by linearity). *By definition* of the Radon-Nikodym derivative, this means that $\mathbb{E}(I_A X) = \mathbb{E}(I_A \mathbb{E}(X|\mathcal{F}))$ for every $A \in \mathcal{F}$, and moreover, by the Radon-Nikodym theorem, there is only one $\mathcal{F}$-measurable random variable $\mathbb{E}(X|\mathcal{F})$ that satisfies this relation. This is precisely Kolmogorov's definition of the conditional expectation.

Definition 2.3.1 (Kolmogorov). Let $X \in \mathcal{L}^1$ and let $\mathcal{F}$ be any σ-algebra. Then $\mathbb{E}(X|\mathcal{F})$ is, by definition, the unique $\mathcal{F}$-measurable random variable that satisfies the relation $\mathbb{E}(I_A X) = \mathbb{E}(I_A \mathbb{E}(X|\mathcal{F}))$ for all events $A \in \mathcal{F}$.

This is usually taken as the starting point in developing conditional expectations, but with your knowledge of martingales it should now be evident that this is just the limiting case of the familiar discrete conditional expectation, but in disguise. On the other hand, this definition is often much easier to deal with: the definition itself does not involve taking any limits (the limits are only involved in proving *existence* of an object that satisfies the definition!)

Using the Kolmogorov definition of the conditional expectation, the following is not hard to prove. We will leave it for you as an exercise.

Theorem 2.3.2 (Elementary properties). *All statements of lemma 2.1.4 still hold in the general case, i.e., when $\mathcal{G}$ and $\mathcal{H}$ are not necessarily finite.*

The following property makes precise the idea that $\mathbb{E}(X|\mathcal{F})$ can be interpreted as the *best estimate* of X given the information $\mathcal{F}$. In other words, we give the conditional expectation (a probabilistic concept) a *statistical* interpretation.

Proposition 2.3.3 (Least squares property). *Let $X \in \mathcal{L}^2$. Then $\mathbb{E}(X|\mathcal{F})$ is the least-mean-square estimate of X given $\mathcal{F}$, i.e., $\mathbb{E}(X|\mathcal{F})$ is the unique $\mathcal{F}$-measurable random variable that satisfies $\mathbb{E}((X - \mathbb{E}(X|\mathcal{F}))^2) = \min_{Y \in \mathcal{L}^2(\mathcal{F})} \mathbb{E}((X - Y)^2)$, where $\mathcal{L}^2(\mathcal{F}) = \{Y \in \mathcal{L}^2 : Y \text{ is } \mathcal{F}\text{-measurable}\}$.*

Beside its statistical interpretation, you can also interpret this result *geometrically*: the conditional expectation $\mathbb{E}(X|\mathcal{F})$ is the *orthogonal projection* of $X \in \mathcal{L}^2$ onto the linear subspace $\mathcal{L}^2(\mathcal{F}) \subset \mathcal{L}^2$ with respect to the inner product $\langle X, Y \rangle = \mathbb{E}(XY)$.

Proof. First, note that $\mathbb{E}((X - Y)^2)$ is finite for any $Y \in \mathcal{L}^2(\mathcal{F})$ by Hölder's inequality. Clearly $\mathbb{E}((X - Y)^2) = \mathbb{E}((X - \mathbb{E}(X|\mathcal{F}) + \mathbb{E}(X|\mathcal{F}) - Y)^2)$. But $\Delta = \mathbb{E}(X|\mathcal{F}) - Y$ is $\mathcal{F}$-measurable by construction, so using the elementary properties of the conditional expectation (and Hölder's inequality to show that $X\Delta \in \mathcal{L}^1$) we see that $\mathbb{E}(\mathbb{E}(X|\mathcal{F})\Delta) = \mathbb{E}(X\Delta)$. Thus $\mathbb{E}((X - Y)^2) = \mathbb{E}((X - \mathbb{E}(X|\mathcal{F})^2) + \mathbb{E}(\Delta^2)$. [Recall the geometric intuition: $\Delta \in \mathcal{L}^2(\mathcal{F})$ and $\mathbb{E}(X|\mathcal{F})$ is the orthogonal projection of X onto $\mathcal{L}^2(\mathcal{F})$, so $H = X - \mathbb{E}(X|\mathcal{F}) \perp \mathcal{L}^2(\mathcal{F})$, and thus $\langle H, \Delta \rangle = \mathbb{E}(H\Delta) = 0$.] The least squares property follows, as $\mathbb{E}(\Delta^2) \geq 0$.

To prove that the minimum is unique, suppose that Y_* is another $\mathcal{F}$-measurable random variable that minimizes $\mathbb{E}((X - Y)^2)$. Then $\mathbb{E}((X - Y_*)^2) = \mathbb{E}((X - \mathbb{E}(X|\mathcal{F}))^2)$. But by

the general formula above, $\mathbb{E}((X - Y_*)^2) = \mathbb{E}((X - \mathbb{E}(X|\mathcal{F}))^2) + \mathbb{E}((\mathbb{E}(X|\mathcal{F}) - Y_*)^2)$. It follows that we must have $\mathbb{E}((\mathbb{E}(X|\mathcal{F}) - Y_*)^2) = 0$; this implies that $\mathbb{E}(X|\mathcal{F}) = Y_*$ a.s. □

Let us briefly remark on the various definitions and constructions of the conditional expectation. We then move on to martingales.

Remark 2.3.4. There are three approaches to defining the conditional expectation.

The first method is Kolmogorov's abstract definition. It is the most difficult to interpret directly, but is the cleanest and usually the easiest to use. Proving uniqueness of the conditional expectation directly using Kolmogorov's definition is easy (do it!), but proving existence is hard—it requires the Radon-Nikodym theorem.

The second method is to define the conditional expectation as the least mean square estimator. This is quite intuitive (certainly from a statistical point of view), and proving existence and uniqueness of $\mathbb{E}(X|\mathcal{F})$ is not difficult provided one first investigates in more detail the geometric properties of the space $\mathcal{L}^2$. However, this definition only works (and is natural) for $X \in \mathcal{L}^2$, so that the conditional expectation has to be extended to $\mathcal{L}^1$ at the end of the day (by approximation).

The third method is the one we used previously, i.e., defining the conditional expectation as a limit of discrete conditional expectations. This is perhaps most intuitive from a probabilistic point of view, but it only seems natural for separable σ-algebras (the extension to the non-separable case being somewhat abstract). Contrary to the previous methods, it is existence that is easy to prove here (using the martingale convergence theorem), but uniqueness is the difficult part.

Kolmogorov's definition of conditional expectations is now universally accepted in probability theory. However, *all* the techniques used above (including geometric and martingale techniques) are very important and are used throughout the subject.

Martingales, supermartingales, submartingales

We have already discussed martingales, and though we have nominally only provided proofs for the discrete case you can easily verify that none of the arguments depended on this; to extend to the general case, just use theorem 2.3.2 instead of lemma 2.1.4. In particular, *the Doob decomposition, the fact that a martingale transform is again a martingle (if it is $\mathcal{L}^1$), and the martingale convergence theorem all hold even if $\mathcal{F}_n$ are not finite (or even separable).* In this section, we will prove some other important properties of martingales and related processes which we have not yet discussed.

First, let us introduce two related types of processes.

Definition 2.3.5. An $\mathcal{F}_n$-adapted stochastic process $\{X_n\}$ is said to be a *supermartingale* if $\mathbb{E}(X_n|\mathcal{F}_m) \leq X_m$ a.s. for every $m \leq n$, and a *submartingale* if $\mathbb{E}(X_n|\mathcal{F}_m) \geq X_m$ a.s. for every $m \leq n$. Hence a *martingale* is a process that is both a supermartingale and a submartingale.

The terminology might be a little confusing at first: a *super*martingale *decreases* on average, while a *sub*martingale *increases* on average. That's how it is.

Example 2.3.6. The winnings in most casinos form a supermartingale.

Example 2.3.7. The price of a stock in simple models of financial markets is often a submartingale (we win on average—otherwise it would not be prudent to invest).

Remark 2.3.8. To prove that a process X_n is a supermartingale, it suffices to check that $\mathbb{E}(X_n|\mathcal{F}_{n-1}) \leq X_{n-1}$ a.s. for all n (why?); similarly, it suffices to check that $\mathbb{E}(X_n|\mathcal{F}_{n-1}) \geq X_{n-1}$ or $\mathbb{E}(X_n|\mathcal{F}_{n-1}) = X_{n-1}$ to demonstrate the submartingale and the martingale property, respectively.

Here are some simple results about supermartingales. You should easily be able to prove these yourself. Do this now. Note that it is often straightforward to extend such results to submartingales by noting that if X_n is an $\mathcal{F}_n$-submartingale, then $K - X_n$ is a supermartingale for any $\mathcal{F}_0$-measurable K.

Lemma 2.3.9. *Let X_n be a supermartingale. Then it can be written uniquely as $X_n = X_0 + A_n + M_n$, where M_n is a martingale and A_n is a nonincreasing predictable process (i.e., $A_n \leq A_{n-1}$ a.s. for all n).*

Lemma 2.3.10. *Let X_n be a supermartingale such that $\sup_n \mathbb{E}(|X_n|) < \infty$. Then there exists a random variable X_∞ such that $X_n \to X_\infty$ a.s.*

Lemma 2.3.11. *Let X_n be a supermartingale and let $A_n \in \mathcal{L}^1$ be a nonnegative predictable process. Then $(A \cdot X)_n$ is a supermartingale, provided it is in $\mathcal{L}^1$.*

Stopping times and optional stopping

A very important notion, in the context of any stochastic process, is a *stopping time*. We will see many of these later on, and, in fact, and entire branch of stochastic control (optimal stopping) is devoted to them! Roughly speaking, an $\mathcal{F}_n$-stopping time is a random time which is "observable" given $\mathcal{F}_n$. In particular, if we know the answer to every yes-no question in $\mathcal{F}_n$, then we also know whether the stopping time has already elapsed (and if so, when) or whether it has yet to happen.

Definition 2.3.12. An ($\mathcal{F}_n$-) stopping time is a random time $\tau : \Omega \to \{0, 1, \ldots, \infty\}$ such that $\{\omega \in \Omega : \tau(\omega) \leq n\} \in \mathcal{F}_n$ for every n.

Example 2.3.13. Let X_n be a stochastic process and $\mathcal{F}_n = \sigma\{X_k : k = 1, \ldots, n\}$. Let $\tau = \inf\{k : X_k \geq 17\}$. Then τ is a stopping time (why?). Note the intuition: if we have observed the process X up to time n, then we can determine from this whether $\tau \leq n$ or $\tau > n$ (after all, if we have been observing the process, then we *know* whether it has already exceeded 17 or not). In the former case, we also know the value of τ (why? prove that $\{\tau = k\} \in \mathcal{F}_n$ for $k \leq n$), but not in the latter case.

Example 2.3.14. Here is an example of a random time that is *not* a stopping time. Recally that a bounded martingale M_n has finitely many upcrossings of the interval $[a, b]$. It would be interesting to study the random time τ at which M_n finishes its *final* upcrossing (i.e., the last time that M_n exceeds b after having previously dipped below a). However, τ is not a stopping time: to know that M_n has up-crossed for the last time, we need to look into the future to determine that it will never up-cross again.

Using the notion of a stopping time, we can define stopped processes.

Definition 2.3.15. Let $\{X_n\}$ be a stochastic process and $\tau < \infty$ be a stopping time. Then X_τ denotes the random variable $X_{\tau(\omega)}(\omega)$: i.e., this is the process X_n evaluated at τ. For any stopping time τ, the stochastic process $X'_n(\omega) = X_{n\wedge\tau(\omega)}(\omega)$ is called the *stopped process*: i.e., $X'_n = X_n$ for $n < \tau$, and $X'_n = X_\tau$ for $n \geq \tau$.

You can consider the stopped process as yet another gambling strategy. Indeed, the process $I_{n\leq\tau}$ is predictable (as $\{\omega : n \leq \tau(\omega)\} = \Omega\backslash\{\omega : \tau(\omega) \leq n-1\} \in \mathcal{F}_{n-1}$), and we can clearly write for any process X_n

$$X_{n\wedge\tau} = X_0 + \sum_{k=1}^{n} I_{k\leq\tau}(X_k - X_{k-1}).$$

Note that this also proves that $X_{n\wedge\tau}$ is again $\mathcal{F}_n$-measurable! (This would not be true if τ were any old random time, rather than a stopping time.)

Speaking of measurability, you might wonder what σ-algebra X_τ is naturally measurable with respect to. This following definition clarifies this point.

Definition 2.3.16. Let $\mathcal{F}_n$ be a filtration and let τ be a stopping time. By definition, $\mathcal{F}_\tau = \{A \in \mathcal{F}_\infty : A \cap \{\tau \leq n\} \in \mathcal{F}_n \text{ for all } n\}$ is the σ-algebra of events that occur before time τ (recall that $\mathcal{F}_\infty = \sigma\{\mathcal{F}_n : n = 1, 2, \ldots\}$). If $\tau < \infty$ a.s., then X_τ is well defined and $\mathcal{F}_\tau$-measurable (why?).

Now suppose that X_n is a martingale (or a super- or submartingale). By the above representation for the stopped process, it is immediately evident that even the stopped process is a martingale (or super- or submartingale, respectively), confirming our intuition that we can not make money on average using a predictable strategy.

Lemma 2.3.17. *If M_n is a martingale (or super-, submartingale) and τ is a stopping time, then $M_{n\wedge\tau}$ is again a martingale (or super-, submartingale, respectively).*

In particular, it follows directly that $\mathbb{E}(M_{n\wedge\tau}) = M_0$ for any n. *However, this does not necessarily imply that* $\mathbb{E}(M_\tau) = 0$, even if $\tau < \infty$ a.s.! To conclude the latter, we need some additional constraints. We will see an example below; if this seems abstract to you, skip ahead to the example.

Theorem 2.3.18 (Optional stopping). *Let M_n be a martingale, and let $\tau < \infty$ be a stopping time. Then $\mathbb{E}(M_\tau) = \mathbb{E}(M_0)$ holds under any of the following conditions: (a) $\tau < K$ a.s. for some $K \in \mathbb{N}$; (b) $|M_n| \leq K$ for some $K \in [0, \infty[$ and all n; (c) $|M_n - M_{n-1}| \leq K$ a.s. for some $K \in [0, \infty[$ and all n, and $\mathbb{E}(\tau) < \infty$. If M_n is a supermartingale, then under the above conditions $\mathbb{E}(M_\tau) \leq \mathbb{E}(M_0)$.*

Remark 2.3.19. There are various extensions of this result; for example, if σ and τ are stopping times and $\sigma \leq \tau$ a.s., then (for example if $\tau \leq K$ a.s.) $\mathbb{E}(M_\tau|\mathcal{F}_\sigma) = M_\sigma$ a.s. We will not need such results, but proving this is good practice!

Proof. We prove the martingale case; the supermartingale result follows identically. To prove (a), it suffices to note that $\mathbb{E}(M_\tau) = \mathbb{E}(M_{K\wedge\tau}) = \mathbb{E}(M_0)$. For (b), note that $M_{n\wedge\tau} \to M_\tau$ a.s. as $n \to \infty$ (by $\tau < \infty$), so the result follows by dominated convergence. For (c), note that

$$|X_{n\wedge\tau}| \le |X_0| + \sum_{k=1}^{n} I_{k\le\tau}|X_k - X_{k-1}| \le |X_0| + K(n\wedge\tau) \le |X_0| + K\tau,$$

where the right hand side is integrable by assumption. Now apply dominated convergence. □

We now give an illuminating example. Make sure you understand this example, and reevaluate what you know about martingales, gambling strategies and fair games.

Example 2.3.20. Let $\xi_1, \xi_2, \ldots$ be independent random variables which take the values ± 1 with equal probability. Define $M_n = M_0 + \sum_{k=1}^{n} \xi_k$. M_n are our winnings in the following fair game: we repeatedly flip a coin; if it comes up heads we gain a dollar, else we lose one. Proving that M_n is a martingale is a piece of cake (do it!)

First, we should note that M_n *a.s. does not converge* as $n \to \infty$. This is practically a trivial observation. If $M_n(\omega)$ were to converge for some path ω, then for sufficiently large N we should have $|M_n(\omega) - M_\infty(\omega)| < \frac{1}{2}$ for all $n \ge N$. But M_n takes only integer values, so this would imply that $M_n(\omega) = K(\omega)$ for some $K(\omega) \in \mathbb{Z}$ and for all $n \ge N$. Such paths clearly have measure zero (as M_n always changes between two time steps: $|M_n - M_{n-1}| = 1$ a.s.) Of course M_n does not satisfy the conditions of the martingale convergence theorem, so we are not surprised.

Now introduce the following stopping time: $\tau = \inf\{n : M_n \ge 2M_0\}$. That is, τ is the first time we have doubled our initial capital. Our strategy is to wait until this happens, then to stop playing, and the question is: do we ever reach this point, i.e., is $\tau < \infty$? Surprisingly, the answer is *yes*! Note that $M_{n\wedge\tau}$ is again a martingale, but $M_{n\wedge\tau} \le 2M_0$ a.s. for all n. Hence this martingale satisfies the condition of the martingale convergence theorem, and so $M_{n\wedge\tau}$ converges as $n \to \infty$. But repeating the argument above, the only way this can happen is if $M_{n\wedge\tau}$ "gets stuck" at $2M_0$—i.e., if $\tau < \infty$ a.s. Apparently *we always double our capital with this strategy!*

We are now in a paradox, and there are several ways out, all of which you should make sure you understand. First, note that $M_\tau = 2M_0$ by construction. Hence $\mathbb{E}(M_\tau) \ne \mathbb{E}(M_0)$, as you would expect. Let us use the optional stopping theorem in reverse. Clearly M_n is a martingale, $\tau < \infty$, and $|M_n - M_{n-1}| \le 1$ for all n. Nonetheless $\mathbb{E}(M_\tau) \ne \mathbb{E}(M_0)$, so evidently $\mathbb{E}(\tau) = \infty$—though we will eventually double our profits, *this will take arbitrarily long on average*. Evidently you can make money on average in a fair game—sometimes—but certainly not on any *finite* time interval! But we already know this, because $\mathbb{E}(M_{n\wedge\tau}) = \mathbb{E}(M_0)$ for any finite n.

Second, note that $M_{n\wedge\tau} \to M_\tau$ a.s, but it is *not* the case that $M_{n\wedge\tau} \to M_\tau$ in $\mathcal{L}^1$; after all, the latter would imply that $\mathbb{E}(M_{n\wedge\tau}) \to \mathbb{E}(M_\tau)$, which we have seen is untrue. But recall when a process does not converge in $\mathcal{L}^1$, our intuition was that the "outliers" of the process somehow grow very rapidly in time. In particular, we have seen that we eventually double our profit, but in the intermediate period we may have to incure *huge* losses in order to keep the game fair.

With the help of the optional sampling theorem, we can actually quantify this idea! What we will do is impose also a *lower* bound on our winnings: once we sink below

a certain value $-R$ (we are R dollars in debt), we go bankrupt and can not continue playing. Our new stopping time is $\kappa = \inf\{n : M_n \geq 2M_0 \text{ or } M_n \leq -R\}$ (κ is the time at which we either reach our target profit, or go bankrupt). Now note that M_κ *does* satisfy the conditions of the optional stopping theorem (as $|M_{n\wedge\kappa}| \leq R \vee 2M_0$), so $\mathbb{E}(M_\kappa) = \mathbb{E}(M_0)$. But M_κ can only take one of two values $-R$ and $2M_0$, so we can explicitly calculate the probability of going bankrupt. For example, if $R = 0$, then we go bankrupt and double our capital with equal probability.

Evidently we can not circumvent our previous conclusion—that no money can be made, on average, in a fair game—*unless we allow ourselves to wait an arbitrarily long time and to go arbitrarily far into debt*. This is closely related to the gambling origin of the word martingale. In 19th century France, various betting strategies were directly based on the idea that if you play long enough, you will make a profit for sure. Such strategies were called martingales, and were firmly believed in by some—until they went bankrupt. In the contemporary words of W. M. Thackeray,

> "You have not played as yet? Do not do so; above all avoid a martingale, if you do. [. . .] I have calculated infallibly, and what has been the effect? Gousset empty, tiroirs empty, necessaire parted for Strasbourg!"
>
> — W. M. Thackeray, *The Newcomes* (1854).

Following the work of Doob we will not follow his advice, but this does not take away from the fact that the original martingale is not recommended as a gambling strategy.

There is much more theory on exactly when martingales converge, and what consequences this has, particularly surrounding the important notion of uniform integrability. We will not cover this here (we want to actually make it to stochastic calculus before the end of term!), but if you wish to do anything probabilistic a good understanding of these topics is indispensable and well worth the effort.

A supermartingale inequality

Let us discuss one more elementary application of martingales and stopping times, which is useful in the study of stochastic stability.

Let M_n be a *nonnegative* supermartingale. By the martingale convergence theorem, $M_n \to M_\infty$ a.s. as $n \to \infty$. Let us now set some threshold $K > 0$; for some K, the limit M_∞ will lie below K with nonzero probability. This does not mean, however, that the sample paths of M_n do not exceed the threshold K before ultimately converging to M_∞, even for those paths where $M_\infty < K$. We could thus ask the question: what is the probability that the sample paths of M_n will never exceed some threshold K? Armed with stopping times, martingale theory, and elementary probability, we can proceed to say something about this question.

Lemma 2.3.21. *Let M_n be an a.s. nonnegative supermartingale and $K > 0$. Then*

$$\mathbb{P}\left(\sup_n M_n \geq K\right) \leq \frac{\mathbb{E}(M_0)}{K}.$$

In particular, for any threshold K, the probability of ever exceeding K can be made arbitrarily small by starting the martingale close to zero.

Proof. Let us first consider a finite time interval, i.e., let us calculate $\mathbb{P}(\sup_{n\le N} M_n \ge K)$ for $N < \infty$. The key is to note that $\{\omega : \sup_{n\le N} M_n(\omega) \ge K\} = \{\omega : M_{\tau(\omega)\wedge N}(\omega) \ge K\}$, where τ is the stopping time $\tau = \inf\{n : M_n \ge K\}$. After all, if $\sup_{n\le N} M_n \ge K$ then $\tau \le N$, so $M_{\tau\wedge N} \ge K$. Conversely, if $\sup_{n\le N} M_n < K$ then $\tau > N$, so $M_{\tau\wedge N} < K$. Using Chebyshev's inequality and the supermartingale property, we have

$$\mathbb{P}\left(\sup_{n\le N} M_n \ge K\right) = \mathbb{P}(M_{\tau\wedge N} \ge K) \le \frac{\mathbb{E}(M_{\tau\wedge N})}{K} \le \frac{\mathbb{E}(M_0)}{K}.$$

Now let $N \to \infty$ using monotone convergence (why monotone?), and we are done. □

Once again, this type of result can be generalized in various ways, and one can also obtain bounds on the moments $\mathbb{E}(\sup_n |M_n|^p)$. You can try to prove these yourself, or look them up in the literature if you need them.

2.4 Some subtleties of continuous time

Up to this point we have only dealt with stochastic processes in discrete time. This course, however, is based on stochastic calculus, which operates exclusively in continuous time. Thus we eventually have to stray into the world of continuous time stochastic processes, and that time has now come.

The theory of continuous time stochastic processes can be much more technical than its discrete time counterpart. Doob's book [Doo53] was one of the first places where such problems were seriously investigated, and the theory was developed over the next 30 or so years into its definitive form, by Doob and his coworkers and particularly by the French probability school of P.-A. Meyer. The ultimate form of the theory—the so-called *théorie générale des processus*—is beautifully developed in the classic books by Dellacherie and Meyer [DM78, DM82] (for a different approach, see [Pro04]). We do *not* want to go this way! The technicalities of the general theory will not be of much help at this level, and will only make our life difficult.

The good news is that we will almost always be able to avoid the problems of continuous time by working with a very special class of continuous time stochastic processes: those with *continuous sample paths*. You can imagine why this would simplify matters: continuous paths are determined by their values on a countable dense set of times (e.g., if we know the values of a continuous function for all rational numbers, then we know the entire function). Once we restrict our attention to countable collections of random variables, many of the technicalities cease to be an issue. We will generally not be too nitpicky about such issues in these notes; the goal of this section is to give you a small glimpse at the issues in continuous time, and to convince you that such issues are not too problematic when we have continuous sample paths.

Equivalent processes and measurability

We will often work with stochastic processes either on a finite time interval $[0, T]$, or on the infinite interval $[0, \infty[$. In either case, a *stochastic process* on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is simply a family $\{X_t\}$ of measurable random variables, indexed by time t. As usual, we are only interested in defining such processes with probability one; but in continuous time, this is a little ambiguous.

Definition 2.4.1. Let X_t and Y_t be two (discrete or continuous time) stochastic processes. Then X_t and Y_t are said to be *indistinguishable* if $\mathbb{P}(X_t = Y_t \text{ for all } t) = 1$, and they are said to be *modifications* of each other if $\mathbb{P}(X_t = Y_t) = 1$ for all t.

Clearly if X_t and Y_t are indistinguishable, then they are modifications (why?). In discrete time, the converse is also true: after all,

$$\mathbb{P}(X_n = Y_n \text{ for all } n) = \mathbb{P}\left(\bigcap_n \{\omega : X_n(\omega) = Y_n(\omega)\}\right) \geq 1 - \sum_n \mathbb{P}(X_n \neq Y_n),$$

so if X_n and Y_n are modifications, then they are indistinguishable. In continuous time, however, the intersection in this expression is uncountable; in the absence of further information (e.g., if we only know that X_t is a modification of Y_t), we can not even show that $\bigcap_t \{\omega : X_t(\omega) = Y_t(\omega)\} \in \mathcal{F}$, let alone that it has unit probability! This requirement is implicit, however, in the definition of indistinguishability.[2]

Example 2.4.2. Let ξ be a Gaussian random variable (with zero mean and unit variance, say), and define the stochastic process $X_t = I_{t<\xi^2}$. Now define $X'_t = I_{t\leq\xi^2}$. For fixed $t > 0$, note that $\mathbb{P}(X_t = X'_t) = \mathbb{P}(\xi^2 \neq t) = 1$. However, $\mathbb{P}(X_t = X'_t \text{ for all } t) = 0$: after all, for every ω, we have $X_t(\omega) \neq X'_t(\omega)$ for $t = \xi(\omega)^2$. Hence X_t and X'_t are modifications, but they are not indistinguishable processes.

One could think that for most practical purposes, processes which are modifications could be considered to be essentially equivalent. Though this intuition is probably justified, we can get into serious trouble by making a modification. What follows are two examples, the first of which is often dealt with by making some assumptions. The second is more serious. Recall that a filtration is a family of σ-algebras $\mathcal{F}_t$, indexed by time, such that $\mathcal{F}_s \subset \mathcal{F}_t$ for all $s \leq t$.

Example 2.4.3. Let ξ be a Gaussian random variable, and consider the processes $X_t = 1$ and $X'_t = I_{t\neq\xi^2}$. Then X_t and X'_t are modifications. Now denote by $\mathcal{F}_t = \sigma\{X_s : s \leq t\}$ the filtration generated by X_t. Then X_t is $\mathcal{F}_t$-adapted, but X'_t is not! Hence evidently modification need not preserve adaptedness.

Though the example is somewhat artificial, it shows that a modification of a stochastic process does not always share all the desirable properties of the process. This particular issue is often suppressed, however, by "augmenting" the σ-algebra $\mathcal{F}_0$ by adding to it all sets of measure zero; this clearly solves the problem, albeit in a way that is arguably as artificial as the problem itself. More serious is the following issue.

Example 2.4.4. Let ξ be a Gaussian random variable, and define $X_t = \xi t$, $X'_t = X_t I_{|X_t|\neq 1}$, and $\mathcal{F}_t = \sigma\{X_s : s \leq t\}$. Then X_t and X'_t are both $\mathcal{F}_t$-adapted and are modifications. Now define the stopping times $\tau = \inf\{t : X_t \geq 1\}$, and similarly $\tau' = \inf\{t : X'_t \geq 1\}$. Note that $\tau = \tau'$! Nonetheless $X_\tau = 1$, while $X'_\tau = 0$.

[2] It would also make sense to require that $\bigcap_t \{\omega : X_t(\omega) = Y_t(\omega)\}$ contains a set of measure one, even if it is not itself measurable. Under certain hypotheses on the probability space (that it be *complete*), this implies that the set is itself measurable. Such details are important if you want to have a full understanding of the mathematics. We will not put a strong emphasis on such issues in this course.

Evidently modification may not preserve the value of the process at a stopping time. This will be a real problem that we have to deal with in the theory of optimal stopping with partial observations: more details will follow in chapter 8.

As already mentioned in remark 1.6.7, we often wish to be able to calculate the time integral of a process X_t, and we still want its expectation to be well defined. To this end, we need the process to be measurable not only with respect to the probability space, but also with respect to time, in which case we can apply Fubini's theorem.

Definition 2.4.5. Let X_t be a stochastic process on some filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, \mathbb{P})$ and time set $\mathbb{T} \subset [0, \infty[$ (e.g., $\mathbb{T} = [0, T]$ or $[0, \infty[$). Then X_t is called *adapted* if X_t is $\mathcal{F}_t$-measurable for all t, is called *measurable* if the random variable $X_\cdot : \mathbb{T} \times \Omega \to \mathbb{R}$ is $\mathcal{B}(\mathbb{T}) \times \mathcal{F}$-measurable, and is called *progressively measurable* if $X_\cdot : [0, t] \cap \mathbb{T} \times \Omega \to \mathbb{R}$ is $\mathcal{B}([0, t] \cap \mathbb{T}) \times \mathcal{F}_t$-measurable for all t.

What do these definitions mean? Adapted you know; measurable means that

$$Y_t = \int_0^t X_s \, ds \quad \text{is well defined and } \mathcal{F}\text{-measurable,}$$

and progressively measurable means that

$$Y_t = \int_0^t X_s \, ds \quad \text{is well defined and } \mathcal{F}_t\text{-measurable;}$$

in particular, progressive measurability guarantees that the process Y_t is adapted.[3] Surely this must be true in any reasonable model! A result of Chung and Doob says that every adapted, measurable process has a progressively measurable modification; we will not need such heavy machinery, though.

Continuous processes

Life becomes much easier if X_t has continuous sample paths: i.e., when the function $t \mapsto X_t(\omega)$ is continuous for every ω. In this case most of the major issues are no longer problematic, and we can basically manipulate such processes in a similar manner as in the discrete time setting. Here is a typical argument.

Lemma 2.4.6. *Let X_t be a stochastic process with continuous paths, and let Y_t be another such process. Then if X_t and Y_t are modifications, then they are indistinguishable. If X_t is $\mathcal{F}_t$-adapted and measurable, then it is $\mathcal{F}_t$-progressively measurable.*

Proof. As X_t and Y_t have continuous paths, it suffices to compare them on a countable dense set: i.e., $\mathbb{P}(X_t = Y_t \text{ for all } t) = \mathbb{P}(X_t = Y_t \text{ for all rational } t)$. But the latter is unity whenever X_t and Y_t are modifications, by the same argument as in the discrete time case.

For the second part, construct a sequence of approximate processes $X^k : [0, t] \times \Omega \to \mathbb{R}$ such that $X^k_t(\omega) = X_t(\omega)$ for all $\omega \in \Omega$ and $t = 0, 2^{-k}, \ldots, 2^{-k}\lfloor 2^k t \rfloor$, and such that the sample paths of X^k are piecewise linear. Then $X^k_s(\omega) \to X_s(\omega)$ as $k \to \infty$ for all ω and $s \in [0, t]$. But it is easy to see that every X^k is $\mathcal{B}([0, t]) \times \mathcal{F}_t$-measurable, and the limit of a sequence of measurable maps is again measurable. The result follows. □

[3] For an example where X_t is adapted and measurable, but Y_t is not adapted, see [Let88, example 2.2].

A natural question to ask is whether the usual limit theorems hold even in continuous time. For example, is it true that if $X_t \geq 0$ a.s. for all $t \in [0, \infty[$, then $\mathbb{E}(\liminf_t X_t) \leq \liminf_t \mathbb{E}(X_t)$ (Fatou's lemma)? This could be a potentially tricky question, as it is not clear that the random variable $\liminf_t X_t$ is even measurable! When we have continuous sample paths, however, we can establish that this is the case. Once this is done, extending the basic convergence theorems is straightforward.

Lemma 2.4.7. *Let the process X_t have continuous sample paths. Then the random variables $\inf_t X_t$, $\sup_t X_t$, $\liminf_t X_t$ and $\limsup_t X_t$ are measurable.*

Proof. As the sample paths of X_t are continuous we have, for example, $\inf_t X_t = \inf_{t \in \mathbf{Q}} X_t$ where $\mathbf{Q}$ are the rational numbers. As these are countable, measurability follows from the countable result (lemma 1.3.3). The same holds for $\sup_t X_t$, $\liminf_t X_t$, and $\limsup_t X_t$. □

We can now establish, for example, Fatou's lemma in continuous time. *The continuous time proofs of the monotone convergence theorem and the dominated convergence theorem follow in the same way.*

Lemma 2.4.8. *Let X_t be an a.s. nonnegative stochastic process with continuous sample paths. Then $\mathbb{E}(\liminf_t X_t) \leq \liminf_t \mathbb{E}(X_t)$. If there is a $Y \in \mathcal{L}^1$ such that $X_t \leq Y$ a.s. for all t, then $\mathbb{E}(\limsup_t X_t) \geq \limsup_t \mathbb{E}(X_t)$.*

Proof. By the previous lemma, $\liminf_t X_t$ is measurable so the statement makes sense. Now suppose that the result does not hold, i.e., $\mathbb{E}(\liminf_t X_t) > \liminf_t \mathbb{E}(X_t)$. Then there exists a sequence of times $t_n \nearrow \infty$ such that $\mathbb{E}(\liminf_t X_t) > \liminf_n \mathbb{E}(X_{t_n})$. But note that $\liminf_n X_{t_n} \geq \liminf_t X_t$ by the definition of the inferior limit, so this would imply $\mathbb{E}(\liminf_n X_{t_n}) > \liminf_n \mathbb{E}(X_{t_n})$. However, X_{t_n} is a discrete time stochastic process, and hence $\mathbb{E}(\liminf_n X_{t_n}) \leq \liminf_n \mathbb{E}(X_{t_n})$ follows from the discrete time version of Fatou's lemma. Thus we have a contradiction. The second part of the result follows similarly. □

With a little more effort, we can also extend the martingale convergence theorem.

Theorem 2.4.9. *Let M_t be martingale, i.e., $\mathbb{E}(M_t | \mathcal{F}_s) = M_s$ a.s. for any $s \leq t$, and assume that M_t has continuous sample paths. If any of the following conditios hold: (a) $\sup_t \mathbb{E}(|M_t|) < \infty$; or (b) $\sup_t \mathbb{E}((M_t)^+) < \infty$; or (c) $\sup_t \mathbb{E}((M_t)^-) < \infty$; then there exists an $\mathcal{F}_\infty$-measurable random variable $M_\infty \in \mathcal{L}^1$ s.t. $M_t \to M_\infty$ a.s.*

Proof. We are done if we can extend Doob's upcrossing lemma to the continuous time case; the proof of the martingale convergence theorem then follows identically.

Let $U_T(a, b)$ denote the number of upcrossings of $a < b$ by M_t in the interval $t \in [0, T]$. Now consider the sequence of times $t_n^k = n2^{-k}T$, and denote by $U_T^k(a, b)$ the number of upcrossings of $a < b$ by the discrete time process $M_{t_n^k}$, $n = 0, \ldots, 2^k$. Note that $M_{t_n^k}$ is a discrete time martingale, so by the upcrossing lemma $\mathbb{E}(U_T^k(a, b)) \leq \mathbb{E}((a - M_T)^+)/(b - a)$. We now claim that $U_T^k(a, b) \nearrow U_T(a, b)$, from which the result follows immediately using monotone convergence. To prove the claim, note that as M_t has continuous sample paths and $[0, T]$ is compact, the sample paths of M_t are uniformly continuous on $[0, T]$. Hence we must have $U_T(a, b) < \infty$, and so $U_T(a, b)(\omega) = U_T^k(a, b)(\omega)$ for $k(\omega)$ sufficiently large. □

In addition to martingale convergence, the optional stopping theorem still holds in the continuous case (you can prove this by approximating the stopping time by a sequence of discrete stopping times), the process $M_{t\wedge\tau}$ is again a martingale and is even progressively measurable if M_t is a martingale and τ is a stopping time, and the supermartingale inequality has an immediate continuous counterpart.

Obviously much has been left unsaid, and the topic of continuous time stochastic processes, even in its most elementary form, deserves at least its own chapter if not an entire course. In the following chapters, however, we will move on to other topics. Hopefully you now have a flavor of the difficulties in continuous time and some ways in which these can be resolved. We will make a minimal fuss over such technical issues in the chapters to come, but if you are ever in doubt you should certainly look up the topic in one of the many textbooks on the subject.

2.5 Further reading

Most probability textbooks define the conditional expectation in the sense of Kolmogorov, and use the Radon-Nikodym theorem to prove its existence. For a development through the orthogonal projection in $\mathcal{L}^2$, see Williams [Wil91] or Kallenberg [Kal97]. The proof of the Radon-Nikodym theorem through martingales is originally due to P.-A. Meyer, and we follow Williams [Wil91].

Martingales in discrete time are treated in any good textbook on probability. See Williams [Wil91], for example, or the classic text by Neveu [Nev75]. The grave omission from this chapter of the theory of uniformly integrable martingales should be emphasized again (you want to study this on your own!) The ultimate reference on martingales in continuous time remains Dellacherie and Meyer [DM82]; another excellent reference is Liptser and Shiryaev [LS01a]. A lively introduction to both the discrete and continuous theory can be found in Pollard [Pol02].

Finally, for the theory of stochastic processes in continuous time, including martingales and related processes, you may consult a variety of excellent textbooks; see Dellacherie and Meyer [DM78, DM82], Rogers and Williams [RW00a], Karatzas and Shreve [KS91], Elliott [Ell82], Protter [Pro04], or Bichteler [Bic02], for example.

CHAPTER 3

The Wiener Process

In the Introduction, we described Brownian motion as the limit of a random walk as the time step and mean square displacement per time step converge to zero. The goal of this chapter is to prove that this limit actually coincides with a well defined stochastic process—the *Wiener process*—and we will study its most important properties. For the reasons discussed in the Introduction, this process will play a fundamental role in the rest of this course, both for its own merits and for its connection with white noise.

3.1 Basic properties and uniqueness

Recall that we think of the Wiener process as the limit as $N \to \infty$, in a suitable sense, of the random walk

$$x_t(N) = \sum_{n=1}^{\lfloor Nt \rfloor} \frac{\xi_n}{\sqrt{N}},$$

where ξ_n are i.i.d. random variables with zero mean and unit variance. That there exists a stochastic process that can be thought of as the limit of $x_t(N)$ is not obvious at all: we will have to construct such a process explicitly, which we will do in section 3.2 (see section 3.4 for further comments). On the other hand, any process that can be thought of as the limit of $x_t(N)$ must necessarily have certain elementary properties. For the time being, let us see how much we can say without proving existence.

The guiding idea of the limit as $N \to \infty$ was the central limit theorem. We could never use this theorem to prove existence of the Wiener process: the central limit theorem does not apply to an uncountable collection of random variables $\{x_t(N) : t \in [0, T]\}$. On the other hand, the central limit theorem completely fixes the limiting distribution at any finite number of times $(x_{t_1}(N), \ldots, x_{t_n}(N))$.

Lemma 3.1.1 (Finite dimensional distributions). *For any finite set of times $t_1 < t_2 < \cdots < t_n$, $n < \infty$, the n-dimensional random variable $(x_{t_1}(N), \ldots, x_{t_n}(N))$ converges in law as $N \to \infty$ to an n-dimensional random variable $(x_{t_1}, \ldots, x_{t_n})$ such that $x_{t_1}, x_{t_2} - x_{t_1}, \ldots, x_{t_n} - x_{t_{n-1}}$ are independent Gaussian random variables with zero mean and variance $t_1, t_2 - t_1, \ldots, t_n - t_{n-1}$, respectively.*

Proof. The increments $x_{t_k}(N) - x_{t_{k-1}}(N)$, $k = 1, \ldots, n$ (choose $t_0 = 0$) are independent for any N, so we may consider the limit in law of each of these increments separately. The result follows immediately from the central limit theorem. □

A different aspect of the Wiener process is the regularity of its sample paths. For increasingly large N, the random walk $x_t(N)$ has increasingly small increments. Hence it is intuitively plausible that in the limit as $N \to \infty$, the limiting process x_t will have continuous sample paths. In fact, this *almost* follows from lemma 3.1.1; to be more precise, the following result holds.

Proposition 3.1.2. *Suppose that we have constructed some stochastic process x_t whose finite dimensional distributions are those of lemma 3.1.1. Then there exists a modification $\tilde{x}_t$ of x_t such that $t \mapsto \tilde{x}_t$ is continuous [Recall that the process $\tilde{x}_t$ is a modification of the process x_t whenever $x_t = \tilde{x}_t$ a.s. for all t.]*

The simplest proof of this result is an almost identical copy of the construction we will use to prove existence of the Wiener process; let us thus postpone the proof of proposition 3.1.2 until the next section.

We have now determined all the finite-dimensional distributions of the Wiener process, and we have established that we may choose its sample paths to be continuous. These are precisely the defining properties of the Wiener process.

Definition 3.1.3. A stochastic process W_t is called a *Wiener process* if

1. the finite dimensional distributions of W_t are those of lemma 3.1.1; and
2. the sample paths of W_t are continuous.

An $\mathbb{R}^n$-valued process $W_t = (W_t^1, \ldots, W_t^n)$ is called an *n-dimensional Wiener process* if $W_t^1, \ldots, W_t^n$ are independent Wiener processes.

In order for this to make sense as a definition, we have to establish at least some form of *uniqueness*—two processes W_t and W_t' which both satisfy the definition should have the same properties! Of course we could never require $W_t = W_t'$ a.s., for the same reason that X and X' being (zero mean, unit variance) Gaussian random variables does not mean $X = X'$ a.s. The appropriate sense of uniqueness is that if W_t and W_t' both satisfy the definition above, then they have the same law.

Proposition 3.1.4 (Uniqueness). *If W_t and W_t' are two Wiener processes, then the $C([0,\infty[)$-valued random variables $W_\cdot, W_\cdot' : \Omega \to C([0,\infty[)$ have the same law.*

Remark 3.1.5 ($C([0,\infty[)$-valued random variables). A $C([0,\infty[)$-valued random variable on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is, by definition, a *measurable* map from Ω to $C([0,\infty[)$. But in order to speak of a measurable map, we have to specify a σ-algebra $\mathcal{C}$ on $C([0,\infty[)$. There are two natural possibilities in this case:

1. Any $x \in C([0,\infty[)$ represents an entire continuous path $x = (x_t : t \in [0,\infty[)$. Define for every time t the evaluation map $\pi_t : C([0,\infty[) \to \mathbb{R}$, $\pi_t(x) = x_t$. It is then natural to set $\mathcal{C} = \sigma\{\pi_t : t \in [0,\infty[\}$.

2. As you might know from a course on functional analysis, the natural topology on $C([0,\infty[)$ is the topology of uniform convergence on compact intervals. We could take $\mathcal{C}$ to be the Borel σ-algebra with respect to this topology.

It turns out that these two definitions for $\mathcal{C}$ coincide: see [RW00a, lemma II.82.3]. So fortunately, what we mean by a $C([0,\infty[)$-valued random variable is unambiguous.

Proof of proposition 3.1.4. Let us first establish that $W_\cdot$ is in fact measurable, and hence a random variable (this follows identically for $W'_\cdot$). By assumption W_t is measurable for every t (as $\{W_t\}$ is assumed to be a stochastic process), so $W_t^{-1}(A) \in \mathcal{F}$ for every $A \in \mathcal{B}(\mathbb{R})$. But $W_t = \pi_t(W_\cdot)$, so $W_\cdot^{-1}(B) \in \mathcal{F}$ for every set $B \in \mathcal{C}$ of the form $B = \pi_t^{-1}(A)$. It remains to note that $\mathcal{C} = \sigma\{\pi_t^{-1}(A) : A \in \mathcal{B}(\mathbb{R}),\ t \in [0,\infty[\}$ by construction, so we find that $W_\cdot^{-1}(\mathcal{C}) = \sigma\{W_t^{-1}(A) : A \in \mathcal{B}(\mathbb{R}),\ t \in [0,\infty[\} \subset \mathcal{F}$. Hence $W_\cdot$ is indeed a $C([0,\infty[)$-valued random variable, and the same holds for $W'_\cdot$.

It remains to show that $W_\cdot$ and $W'_\cdot$ have the same law, i.e., that they induce the same probability measure on $(C([0,\infty[), \mathcal{C})$. The usual way to show that two measures coincide is using Dynkin's π-system lemma 1.7.3, and this is indeed what we will do! A *cylinder set* is a set $C \in \mathcal{C}$ of the form $C = \pi_{t_1}^{-1}(A_1) \cap \pi_{t_2}^{-1}(A_2) \cap \cdots \cap \pi_{t_n}^{-1}(A_n)$ for an arbitrary finite number of times $t_1, \ldots, t_n \in [0,\infty[$ and Borel sets $A_1, \ldots, A_n \in \mathcal{B}(\mathbb{R})$. Denote by $\mathcal{C}_{\mathrm{cyl}}$ the collection of all cylinder sets, and note that $\mathcal{C}_{\mathrm{cyl}}$ is a π-system and $\sigma\{\mathcal{C}_{\mathrm{cyl}}\} = \mathcal{C}$. But the definition of the Wiener process specifies completely all finite dimensional distributions, so the laws of any two Wiener processes must coincide on $\mathcal{C}_{\mathrm{cyl}}$. Dynkin's π-system lemma does the rest. □

Given a Wiener process W_t, we can introduce its natural filtration $\mathcal{F}_t^W = \sigma\{W_s : s \le t\}$. More generally, it is sometimes convenient to speak of an $\mathcal{F}_t$-Wiener process.

Definition 3.1.6. Let $\mathcal{F}_t$ be a filtration. Then a stochastic process W_t is called an *$\mathcal{F}_t$-Wiener process* if W_t is a Wiener process, is $\mathcal{F}_t$-adapted, and $W_t - W_s$ is independent of $\mathcal{F}_s$ for any $t > s$. [Note that any Wiener process W_t is an $\mathcal{F}_t^W$-Wiener process.]

Lemma 3.1.7. *An $\mathcal{F}_t$-Wiener process W_t is an $\mathcal{F}_t$-martingale.*

Proof. We need to prove that $\mathbb{E}(W_t|\mathcal{F}_s) = W_s$ for any $t > s$. But as W_s is $\mathcal{F}_s$-measurable (by adaptedness) this is equivalent to $\mathbb{E}(W_t - W_s|\mathcal{F}_s) = 0$, and this is clearly true by the definition of the Wiener process (as $W_t - W_s$ has zero mean and is independent of $\mathcal{F}_s$). □

The Wiener process is also a *Markov process*. You know what this means in discrete time from previous courses, but we have not yet introduced the continuous time definition. Let us do this now.

Definition 3.1.8. An $\mathcal{F}_t$-adapted process X_t is called an *$\mathcal{F}_t$-Markov process* if we have $\mathbb{E}(f(X_t)|\mathcal{F}_s) = \mathbb{E}(f(X_t)|X_s)$ for all $t \ge s$ and all bounded measurable functions f. When the filtration is not specified, the natural filtration $\mathcal{F}_t^X$ is implied.

Lemma 3.1.9. *An $\mathcal{F}_t$-Wiener process W_t is an $\mathcal{F}_t$-Markov process.*

Proof. We have to prove that $\mathbb{E}(f(W_t)|\mathcal{F}_s) = \mathbb{E}(f(W_t)|W_s)$. Note that we can trivially write $f(W_t) = f((W_t - W_s) + W_s)$, where $W_t - W_s$ is independent of $\mathcal{F}_s$ and W_s is $\mathcal{F}_s$-measurable. We claim that $\mathbb{E}(f(W_t)|\mathcal{F}_s) = g(W_s)$ with $g(x) = \mathbb{E}(f(W_t - W_s + x))$. As $g(W_s)$ is $\sigma(W_s)$-measurable, we can then write $\mathbb{E}(f(W_t)|W_s) = \mathbb{E}(\mathbb{E}(f(W_t)|\mathcal{F}_s)|W_s) = \mathbb{E}(g(W_s)|W_s) = g(W_s)$, where we have used $\sigma(W_s) \subset \mathcal{F}_s$. The result now follows.

It remains to prove $\mathbb{E}(f(W_t)|\mathcal{F}_s) = g(W_s)$, or equivalently $\mathbb{E}(g(W_s)I_A) = \mathbb{E}(f(W_t)I_A)$ for all $A \in \mathcal{F}_s$ (by Kolmogorov's definition of the conditional expectation). Consider the pair of random variables $X = W_t - W_s$ and $Y = (W_s, I_A)$, and note that X and Y are independent. Hence by theorem 1.6.6, the law of (X, Y) is a product measure $\mu_X \times \mu_Y$, and so

$$\mathbb{E}(f(W_t)I_A) = \int f(x+w)a\,\mu_X(dx) \times \mu_Y(dw, da) =$$
$$\int \left[\int f(x+w)\,\mu_X(dx) \right] a\,\mu_Y(dw, da) = \int g(w)a\,\mu_Y(dw, da) = \mathbb{E}(g(W_s)I_A).$$

using Fubini's theorem (which applies by the boundedness of f). We are done. □

To complete our discussion of the elementary properties of the Wiener process, let us exhibit some odd properties of its sample paths. The sample paths of the Wiener process are *extremely* irregular, and the study of their properties remains an active topic to this day (see, e.g., [MP06]). We will only consider those properties which we will need to make sense of later developments.

Lemma 3.1.10. *With unit probability, the sample paths of a Wiener process W_t are non-differentiable at any rational time t.*

Proof. Suppose that W_t is differentiable at some point t. Then $\lim_{h \searrow 0}(W_{t+h} - W_t)/h$ exists and is finite, and in particular, there exists a constant $M < \infty$ (depending on ω) such that $|W_{t+h} - W_t|/h < M$ for sufficiently small $h > 0$. We will show that with unit probability this cannot be true. Set $h = n^{-1}$ where n is integer; then $|W_{t+h} - W_t|/h < M$ for sufficiently small $h > 0$ implies that $\sup_{n \geq 1} n|W_{t+n^{-1}}(\omega) - W_t(\omega)| < \infty$. But we can write (why?)

$$\left\{ \omega : \sup_{n \geq 1} n|W_{t+n^{-1}}(\omega) - W_t(\omega)| < \infty \right\} = \bigcup_{M \geq 1} \bigcap_{n \geq 1} \{\omega : n|W_{t+n^{-1}}(\omega) - W_t(\omega)| < M\}.$$

Using simple set manipulations, we obtain

$$\mathbb{P}\left(\sup_{n \geq 1} n|W_{t+n^{-1}} - W_t| < \infty \right) \leq \lim_{M \to \infty} \inf_{n \geq 1} \mathbb{P}(n|W_{t+n^{-1}} - W_t| < M).$$

But $W_{t+n^{-1}} - W_t$ is a Gaussian random variable with zero mean and variance n^{-1}, so

$$\inf_{n \geq 1} \mathbb{P}(n|W_{t+n^{-1}} - W_t| < M) = \inf_{n \geq 1} \mathbb{P}(|\xi| < Mn^{-1/2}) = 0,$$

where ξ is a canonical Gaussian random variable with zero mean and unit variance. Hence we find that $\mathbb{P}(\lim_{n^{-1} \searrow 0}(W_{t+n^{-1}} - W_t)/n^{-1}$ is finite$) = 0$, so W_t is a.s. not differentiable at t for any fixed time t. But as the rational numbers are countable, the result follows. □

Apparently the sample paths of Brownian motion are very rough; certainly the derivative of the Wiener process cannot be a sensible stochastic process, once again confirming the fact that white noise is not a stochastic process (compare with the

discussion in the Introduction). With a little more work one can show that with unit probability, the sample paths of the Wiener process are not differentiable at *any* time t (this does not follow trivially from the previous result, as the set of all times t is not countable); see [KS91, theorem 2.9.18] for a proof.

Another measure of the irregularity of the sample paths of the Wiener process is their total variation. For any real-valued function $f(t)$, the *total variation* of f on the interval $t \in [a,b]$ is defined as

$$\mathrm{TV}(f,a,b) = \sup_{k\geq 0} \sup_{(t_i)\in P(k,a,b)} \sum_{i=0}^{k} |f(t_{i+1}) - f(t_i)|,$$

where $P(k,a,b)$ denotes the set of all partitions $a = t_0 < t_1 < \cdots < t_k < t_{k+1} = b$. You can think of the total variation as follows: suppose that we are driving around in a car, and $f(t)$ denotes our position at time t. Then $\mathrm{TV}(f,a,b)$ is the total distance which we have travelled in the time interval $[a,b]$ (i.e., if our car were to go for a fixed number of miles per gallon, then $\mathrm{TV}(f,a,b)$ would be the amount of fuel which we used up between time a and time b). Note that even when $\sup_{a\leq s\leq t\leq b} |f(t) - f(s)|$ is small, we could still travel a significant total distance if we oscillate very rapidly in the interval $[a,b]$. But the Wiener process, whose time derivative is infinite at every (rational) time, must oscillate very rapidly indeed!

Lemma 3.1.11. *With unit probability,* $\mathrm{TV}(W_\cdot,a,b) = \infty$ *for any* $a < b$*. In other words, the sample paths of the Wiener process are a.s. of infinite variation.*

Proof. Denote by $P(a,b) = \bigcup_{k\geq 0} P(k,a,b)$ the set of all finite partitions of $[a,b]$. We are done if we can find a sequence of partitions $\pi_n \in P(a,b)$ such that

$$\sum_{t_i\in\pi_n} |W_{t_{i+1}} - W_{t_i}| \xrightarrow{n\to\infty} \infty \quad \text{a.s.}$$

To this end, let us concentrate on a slightly different object. For any $\pi \in P(a,b)$, we have

$$\mathbb{E} \sum_{t_i\in\pi} (W_{t_{i+1}} - W_{t_i})^2 = \sum_{t_i\in\pi} (t_{i+1} - t_i) = b - a.$$

Call $Z_i = (W_{t_{i+1}} - W_{t_i})^2 - (t_{i+1} - t_i)$, and note that for different i, the random variables Z_i are independent and have the same law as $(\xi^2 - 1)(t_{i+1} - t_i)$, where ξ is a Gaussian random variable with zero mean and unit variance. Hence

$$\mathbb{E}\left[\left(\sum_{t_i\in\pi} (W_{t_{i+1}} - W_{t_i})^2 - (b-a)\right)^2\right] = \mathbb{E}\left[\sum_{t_i\in\pi} Z_i^2\right] = \mathbb{E}((\xi^2-1)^2) \sum_{t_i\in\pi} (t_{i+1} - t_i)^2.$$

Let us now choose a sequence π_n such that $\sup_{t_i\in\pi_n} |t_{i+1} - t_i| \to 0$. Then

$$\mathbb{E}\left[\left(\sum_{t_i\in\pi_n} (W_{t_{i+1}} - W_{t_i})^2 - (b-a)\right)^2\right] \leq (b-a)\mathbb{E}((\xi^2-1)^2) \sup_{t_i\in\pi_n} |t_{i+1} - t_i| \xrightarrow{n\to\infty} 0.$$

In particular, we find that

$$Q_n = \sum_{t_i\in\pi_n} (W_{t_{i+1}} - W_{t_i})^2 \xrightarrow{n\to\infty} b - a \quad \text{in } \mathcal{L}^2,$$

so $Q_n \to b-a$ in probability also, and hence we can find a subsequence $m(n) \nearrow \infty$ such that $Q_{m(n)} \to b-a$ a.s. But then $\mathrm{TV}(W_\cdot, a, b) < \infty$ with nonzero probability would imply

$$b-a \le \lim_{n\to\infty} \left[\left\{ \sup_{t_i \in \pi_{m(n)}} |W_{t_{i+1}} - W_{t_i}| \right\} \sum_{t_i \in \pi_{m(n)}} |W_{t_{i+1}} - W_{t_i}| \right] = 0$$

with nonzero probability (as $\sup_{t_i \in \pi_{m(n)}} |W_{t_{i+1}} - W_{t_i}| \to 0$ by continuity of the sample paths), which contradicts $a < b$. Hence $\mathrm{TV}(W_\cdot, a, b) = \infty$ a.s. for fixed $a < b$. It remains to note that it suffices to consider rational $a < b$; after all, if $\mathrm{TV}(W_\cdot, a, b)$ is finite for some $a < b$, it must be finite for all subintervals of $[a, b]$ with rational endpoints. As we have shown that with unit probability this cannot happen, the proof is complete. □

You might argue that the Wiener process is a very bad model for Brownian motion, as clearly no physical particle can travel an infinite distance in a finite time! But as usual, the Wiener process should be interpreted as an extremely convenient mathematical idealization. Indeed, any physical particle in a fluid will have travelled a humongous total distance, due to the constant bombardment by the fluid molecules, in its diffusion between two (not necessarily distant) points. We have idealized matters by making the total distance truly infinite, but look what we have gained: the martingale property, the Markov property, etc., etc., etc.

It is also the infinite variation property, however, that will get us in trouble when we define stochastic integrals. Recall from the Introduction that we ultimately wish to give meaning to formal integrals of the form $\int_0^t f_s\, \xi_s\, ds$, where ξ_s is white noise, by defining a suitable stochastic integral of the form $\int_0^t f_s\, dW_s$.

The usual way to define such objects is through the *Stieltjes integral*. Forgetting about probability theory for the moment, recall that we write by definition

$$\int_0^t f(s)\, dg(s) = \lim_{\pi} \sum_{t_i \in \pi} f(s_i)\, (g(t_{i+1}) - g(t_i)),$$

where the limit is taken over a sequence of refining partitions $\pi \in P(0,t)$ such that $\max |t_{i+1} - t_i| \to 0$, and s_i is an arbitrary point between t_{i+1} and t_i. When does the limit exist? Well, note that for $\pi_m \subset \pi_n$ (π_n is a finer partition than π_m),

$$\sum_{t_i \in \pi_n} f(s_i)\, (g(t_{i+1}) - g(t_i)) - \sum_{t'_i \in \pi_m} f(s'_i)\, (g(t'_{i+1}) - g(t'_i))$$
$$= \sum_{t_i \in \pi_n} (f(s_i) - f(s''_i))\, (g(t_{i+1}) - g(t_i)),$$

where s''_i is chosen in the obvious way. But note that

$$\left| \sum_{t_i \in \pi_n} (f(s_i) - f(s''_i))\, (g(t_{i+1}) - g(t_i)) \right| \le \mathrm{TV}(g, 0, t) \max_i |f(s_i) - f(s''_i)|.$$

Hence if f is continuous and g is of finite total variation, then the sequence of sums obtained from a refining sequence of partitions π_n is a Cauchy sequence and thus

converges to a unique limit (which we call the Stieltjes integral). More disturbingly, however, one can also prove the converse: *if g is of infinite variation, then there exists a continuous function f such that the Stieltjes integral does not exist.* The proof requires a little functional analysis, so we will not do it here; see [Pro04, section I.8].

The unfortunate conclusion of the story is that the integral $\int_0^t f_s\,dW_s$ cannot be defined in the usual way—at least not if we insist that we can integrate at least continuous processes f_t, which is surely desirable. With a little insight and some amount of work, we will succeed in circumventing this problem. In fact, you can get a hint on how to proceed from the proof of lemma 3.1.11: even though the total variation of the Wiener process is a.s. infinite, the *quadratic variation*

$$\lim_{n\to\infty} \sum_{t_i\in\pi_{m(n)}} (W_{t_{i+1}} - W_{t_i})^2 = b - a \quad \text{a.s.}$$

is finite. Maybe if we square things, things will get better? Indeed they will, though we still need to introduce a crucial insight in order not to violate the impossibility of defining the Stieltjes integral. But we will go into this extensively in the next chapter.

3.2 Existence: a multiscale construction

We are finally ready to construct a Wiener process. In principle, we will do this exactly as one would think: by taking the limit of a sequence of random walks. It is not so easy, however, to prove directly that any random walk of the form $x_t(N)$ defined previously converges to a Wiener process; in what sense should the convergence even be interpreted? Some comments on this matter can be found in section 3.4. Instead, we will concentrate in this section on a *particular* random walk for which convergence is particularly easy to prove. As long as we can verify that the limiting process satisfies definition 3.1.3, we are then done—uniqueness guarantees that there are no other Wiener processes, so to speak.

We will make two straightforward simplifications and one more inspired simplification to our canonical random walk model $x_t(N)$. First, note that we can restrict ourselves to a fixed time interval, say, $t \in [0,1]$. Once we have defined a stochastic process which satisfies definition 3.1.3 for $t \in [0,1]$, we can easily extend it to all of $[0,\infty[$: recall that the increments of the Wiener process are independent!

Lemma 3.2.1. *Let $\{W_t : t \in [0,1]\}$ be a stochastic process on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ that satisfies definition 3.1.3. Then there exists a stochastic process $\{W'_t : t \in [0,\infty[\}$ on a probability space $(\Omega', \mathcal{F}', \mathbb{P}')$ that satisfies definition 3.1.3 for all t.*

Proof. Set[1] $\Omega' = \Omega \times \Omega \times \cdots$, $\mathcal{F}' = \mathcal{F} \times \mathcal{F} \times \cdots$, and $\mathbb{P}' = \mathbb{P} \times \mathbb{P} \times \cdots$. Then Ω' carries an i.i.d. sequence of processes $\{W^n_t, t \in [0,1]\}$, $n = 1, 2, \ldots$. You can easily verify that $W'_t = \sum_{k=1}^{\lfloor t\rfloor} W^k_1 + W^{\lfloor t\rfloor+1}_{t-\lfloor t\rfloor}$ satisfies definition 3.1.3 for all $t \in [0,\infty[$. □

[1] We are cheating a little, as we only defined the countable product space in Theorem 1.6.8 for $\Omega = \mathbb{R}$. In fact, such a space always exists, see [Kal97, corollary 5.18], but for our purposes Theorem 1.6.8 will turn out to be sufficient: every Ω will be itself a space that carries a sequence of independent random variables, and, as in the proof of Theorem 1.6.8, we can always construct such a sequence on $\Omega = \mathbb{R}$.

The second simplification is to make our random walks have continuous sample paths, unlike $x_t(N)$ which has jumps. The reason for this is that there is a very simple way to prove that a sequence of continuous functions converges to a continuous function: this is always the case when the sequence converges *uniformly*. This is an elementary result from calculus, but let us recall it to refresh your memory.

Lemma 3.2.2. *Let $f_n(t)$, $n = 1, 2, \ldots$ be a sequence of continuous functions on $t \in [0, 1]$ that converge uniformly to some function $f(t)$, i.e., $\sup_{t\in[0,1]} |f_n(t) - f(t)| \to 0$ as $n \to \infty$. Then $f(t)$ must be a continuous function.*

Proof. Clearly $|f(x)-f(y)| \le |f(x)-f_n(x)|+|f_n(x)-f_n(y)|+|f_n(y)-f(y)|$ for any n and $x, y \in [0, 1]$. Let $\varepsilon > 0$. Then we can choose n sufficiently large so that $|f_n(x) - f(x)| < \varepsilon/3$ for any x, and then $|f(x) - f(y)| \le 2\varepsilon/3 + |f_n(x) - f_n(y)|$ for all $x, y \in [0, 1]$. But as f_n is uniformly continuous (as $[0, 1]$ is compact), there is a $\delta > 0$ such that $|f_n(x) - f_n(y)| < \varepsilon/3$ for all $|x - y| < \delta$. Hence f satisfies the (ε-δ) definition of continuity. □

Here is another useful trick from the same chapter of your calculus textbook.

Lemma 3.2.3. *Let $f_n(t)$, $n = 1, 2, \ldots$ be a sequence of continuous functions on $t \in [0, 1]$, such that $\sum_n \sup_{t\in[0,1]} |f_{n+1}(t) - f_n(t)| < \infty$. Then $f_n(t)$ converge uniformly to some continuous function $f(t)$.*

Proof. Note that $f_n(t) = f_1(t) + \sum_{k=1}^{n-1}(f_{k+1}(t) - f_k(t))$. By our assumption, the sum is absolutely convergent so $f_n(t) \to f(t)$ as $n \to \infty$ for every $t \in [0, 1]$. It remains to show that the convergence is uniform. But this follows from $\sup_{t\in[0,1]} |f_m(t) - f(t)| = \sup_{t\in[0,1]} |\sum_{k=m}^{\infty}(f_{k+1}(t) - f_k(t))| \le \sum_{k=m}^{\infty} \sup_{t\in[0,1]} |(f_{k+1}(t) - f_k(t))| \to 0$. □

You are probably starting to get a picture of the strategy which we will follow: we will define a sequence W_t^n of random walks with *continuous* sample paths, and attempt to prove that $\sum_n \sup_{t\in[0,1]} |W_t^n - W_t^{n+1}| < \infty$ a.s. We are then guaranteed that W_t^n converges a.s. to some stochastic process W_t with continuous sample paths, and all that remains is to verify the finite dimensional distributions of W_t. But the finite dimensional distributions are the easy part—see, e.g., lemma 3.1.1!

It is at this point, however, that we need a little bit of real insight. Following our intended strategy, it may seem initially that we could define W_t^n just like $x_t(n)$, except that we make the sample paths piecewise linear rather than piecewise constant (e.g., set $W^n_{k/2^n} = \sum_{\ell=1}^{k} \xi_\ell/2^{n/2}$ for $k = 0, \ldots, 2^n$, and interpolate linearly in the time intervals $k/2^n < t < (k + 1)/2^n$). However, this way W_t^n can never converge a.s. as $n \to \infty$. In going from W_t^n to W_t^{n+1}, the process W_t^n gets compressed to the interval $[0, \frac{1}{2}]$, while the increments of W_t^{n+1} on $]\frac{1}{2}, 1]$ are defined using a set of independent random variables $\xi_{2^n+1}, \ldots, \xi_{2^{n+1}}$. This is illustrated in figure 3.1.

Remark 3.2.4. Of course, we do not necessarily expect our random walks to converge almost surely; for example, lemma 3.1.1 was based on the central limit theorem, which only gives convergence in law. The theory of weak convergence, which defines the appropriate notion of convergence in law for stochastic processes, can indeed be used to prove existence of the Wiener process; see [Bil99] or [KS91, section 2.4]. The technicalities involved are a highly nontrivial, however, and we would have to spend

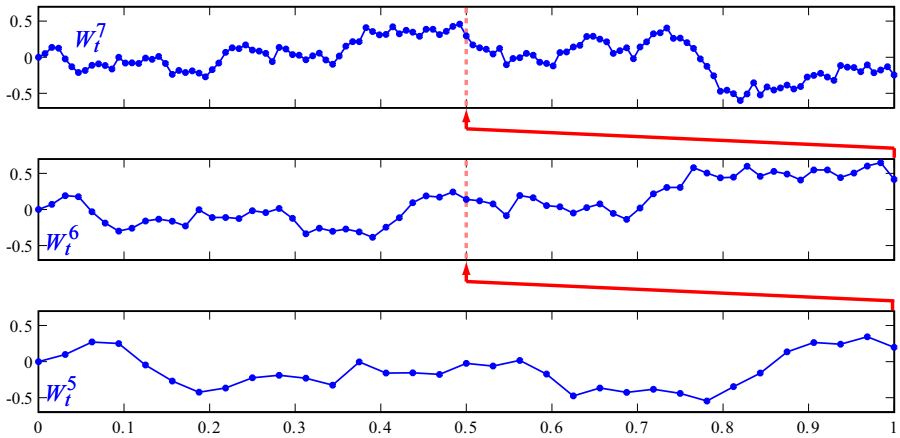

Figure 3.1. Sample paths, given a single realization of $\{\xi_n\}$, of W_t^n for $n = 5, 6, 7$. When n is increased by one, the previous sample path is compressed to the interval $[0, 1/2]$, scaled by $2^{-1/2}$, and the increments in $]1/2, 1]$ are generated from the next batch of independent ξ_ns.

an entire chapter just introducing all the necessary machinery! Instead, we will use a clever trick (due to P. Lévy and Z. Ciesielski) to define a very special sequence of random walks W_t^n which actually converges almost surely. Once we have a.s. convergence, the proofs become much more elementary and intuitive.

The idea is illustrated in figure 3.2. The random walk W_t^n consists 2^n points connected by straight lines. In going from W_t^n to W_t^{n+1}, our previous strategy was to concatenate another 2^n points at the end of the path, and then to compress this new path to fit in the interval $[0, 1]$. Rather than add points at the end of the path, however, we will now add our new 2^n points *in between* the existing nodes of the sample path. This way the shape of the path remains fixed, and we just keep adding detail at finer and finer scales. Then we would certainly expect the sample paths to converge almost surely; the question is whether we can add points between the existing nodes in such a way that the random walks W_t^n have the desired statistics.

Let us work out how to do this. The random walk W_t^n has nodes at $t = k2^{-n}$, $k = 0, \ldots, 2^n$, connected by straight lines. For any further W_t^m with $m > n$, we want to only add nodes between the times $k2^{-n}$, i.e., $W_t^m = W_t^n$ for $t = k2^{-n}$, $k = 0, \ldots, 2^n$. Hence the points $W^n_{k2^{-n}}$ must already be distributed according to the corresponding finite dimensional distribution of lemma 3.1.1: $W^n_{k2^{-n}}$ must be a Gaussian random variable with zero mean and variance $k2^{-n}$, and $W^n_{(k+1)2^{-n}} - W^n_{k2^{-n}}$ must be independent of $W^n_{k2^{-n}}$ for any k. Suppose that we have constructed such a W_t^n. We need to show how to generate W_t^{n+1} from it, by adding points between the existing nodes only, so that W_t^{n+1} has the correct distribution.

Fix n and k, and assume we are given W_t^{n-1}. Let us write

$$Y_0 = W^n_{k2^{-(n-1)}} = W^{n-1}_{k2^{-(n-1)}}, \qquad Y_1 = W^n_{(k+1)2^{-(n-1)}} = W^{n-1}_{(k+1)2^{-(n-1)}}.$$

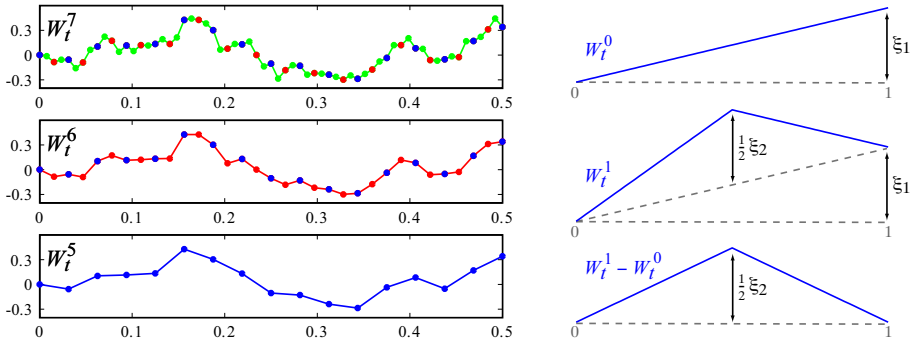

Figure 3.2. Rather than concatenate additional ξ_ns at the end of the sample paths, we define W_t^n for increasing n by adding new ξ_n in between the existing nodes of the sample path. This way detail is added at increasingly fine scales, and the sample path will in fact converge a.s. The procedure is illustrated on the right for W_t^0 and W_t^1; the first path is a line between $W_0^0 = 0$ and $W_1^0 = \xi_1$, while for W_t^1 another point is added at $t = 1/2$ (see text).

We wish to choose a new point $X = W^n_{(2k+1)2^{-n}}$ between Y_0 and Y_1, such that

1. $Y_1 - X$ and $X - Y_0$ are Gaussian with mean zero and variance 2^{-n},
2. $Y_1 - X$, $X - Y_0$ and Y_0 are independent.

We already know the following properties of Y_0 and Y_1:

1. $Y_1 - Y_0$ is Gaussian with mean zero and variance $2^{-(n-1)}$,
2. $Y_1 - Y_0$ and Y_0 are independent.

We claim that $X = (Y_0+Y_1)/2+2^{-(n+1)/2}\xi$, where ξ is a Gaussian random variable with zero mean and unit variance independent of Y_0 and Y_1, satisfies the requirements. Indeed, you can easily verify that $X - Y_0$ and $Y_1 - X$ have the correct mean and variance and are independent of Y_0 (why?). It remains to show that $Y_1 - X$ is independent of $X - Y_0$; but this follows immediately from the well known fact that if ξ_1, ξ_2 are i.i.d. Gaussian, then $\xi_1 + \xi_2$ and $\xi_1 - \xi_2$ are also i.i.d. Gaussian.[2]

How then do we define all of W_t^n from W_t^{n-1}? As is illustrated in the right pane of figure 3.2 for $n = 1$, we need to add to W_t^{n-1} a collection of tent-shaped functions (*Schauder functions*), centered between the nodes of W_t^{n-1} and zero on those nodes, that "lift" the points in between the nodes of W_t^{n-1} by independent random quantities which are Gaussian distributed with zero mean and variance $2^{-(n+1)}$. The best way to see this is to have a good look at figure 3.2; once you have made sure you understand what is going on, we can get down to business.

[2] This follows from the fact that the distribution of an i.i.d. Gaussian random vector $\mathbf{x}$ with zero mean is *isotropic*—it is invariant under orthogonal transformations, i.e., $\mathbf{x}$ has the same law as $A\mathbf{x}$ for any orthogonal matrix A. If you did not know this, now is a good time to prove it!

Theorem 3.2.5. *There exists a Wiener process W_t on some probab. space $(\Omega, \mathcal{F}, \mathbb{P})$.*

Proof. Let us begin by introducing the Schauder (tent-shaped) functions. Note that the derivative of a Schauder function should be piecewise constant: it has a positive value on the increasing slope, a negative value on the decreasing slope, and is zero elsewhere. Such functions are called the *Haar wavelets* $H_{n,k}(t)$ with $n = 0, 1, \ldots$ and $k = 1, 3, 5, \ldots, 2^n - 1$, defined as

$$H_{0,1}(t) = 1, \qquad H_{n,k}(t) = \begin{cases} +2^{(n-1)/2} & (k-1)2^{-n} < t \le k2^{-n}, \\ -2^{(n-1)/2} & k2^{-n} < t \le (k+1)2^{-n}, \\ 0 & \text{otherwise}, \end{cases} \qquad (n \ge 1).$$

The Haar wavelets are localized on increasingly fine length scales for increasing n, while the index k shifts the wavelet across the interval $[0, 1]$. We now define the Schauder functions $S_{n,k}(t)$ simply as indefinite integrals of the Haar wavelets:

$$S_{n,k}(t) = \int_0^t H_{n,k}(s)\, ds, \qquad n = 0, 1, \ldots, \quad k = 1, 3, 5, \ldots, 2^n - 1.$$

You can easily convince yourself that these are precisely the desired tent-shaped functions.

Let us now construct our random walks on $[0, 1]$. Let $(\Omega', \mathcal{F}', \mathbb{P}')$ be a probability space that carries a double sequence $\{\xi_{n,k} : n = 0, 1, \ldots,\ k = 1, 3, \ldots, 2^n - 1\}$ of i.i.d. Gaussian random variables with zero mean and unit variance (which exists by theorem 1.6.8). Figure 3.2 shows how to proceed: clearly $W_t^0 = \xi_{0,1} S_{0,1}(t)$, while $W_t^1 = \xi_{0,1} S_{0,1}(t) + \xi_{1,1} S_{1,1}(t)$ (note the convenient normalization—the tent function $S_{n,k}$ has height $2^{-(n+1)/2}$). Continuing in the same manner, convince yourself that the Nth random walk can be written as

$$W_t^N = \sum_{n=0}^{N} \sum_{k=1,3,\ldots,2^n-1} \xi_{n,k} S_{n,k}(t).$$

We now arrive in the second step of our program: we would like to show that the sequence of processes W_t^N converges uniformly with unit probability, in which case we can define a stochastic process $W_t = \lim_{N\to\infty} W_t^N$ with a.s. continuous sample paths.

We need some simple estimates. First, note that

$$\mathbb{P}\left(\sup_{t\in[0,1]} |W_t^n - W_t^{n-1}| > \varepsilon_n \right) = \mathbb{P}\left(\sup_{k=1,3,\ldots,2^n-1} |\xi_{n,k}| > 2^{(n+1)/2} \varepsilon_n \right),$$

as you can check directly. But note that we can estimate (why?)

$$\mathbb{P}\left(\sup_{k=1,3,\ldots,2^n-1} |\xi_{n,k}| > 2^{(n+1)/2} \varepsilon_n \right) \le \sum_{k=1,3,\ldots,2^n-1} \mathbb{P}(|\xi_{n,k}| > 2^{(n+1)/2} \varepsilon_n),$$

so we can write (using the fact that $\xi_{n,k}$ are i.i.d.)

$$\mathbb{P}\left(\sup_{t\in[0,1]} |W_t^n - W_t^{n-1}| > \varepsilon_n \right) \le 2^{n-1} \mathbb{P}(|\xi_{0,1}| > 2^{(n+1)/2} \varepsilon_n).$$

We need to estimate the latter term, but (as will become evident shortly) a direct application of Chebyshev's inequality is too crude. Instead, let us apply the following trick, which is often useful. Note that $\xi_{1,0}$ is symmetrically distributed around zero, so we have $\mathbb{P}(|\xi_{0,1}| > \alpha) = \mathbb{P}(\xi_{0,1} > \alpha) + \mathbb{P}(\xi_{0,1} < -\alpha) = 2\mathbb{P}(\xi_{0,1} > \alpha)$. Now use Chebyshev's inequality as follows:

$$\mathbb{P}(\xi_{0,1} > \alpha) = \mathbb{P}(e^{\xi_{0,1}} > e^{\alpha}) \le e^{-\alpha} \mathbb{E}(e^{\xi_{0,1}}) = e^{1/2-\alpha}.$$

We thus obtain

$$\mathbb{P}\left(\sup_{t\in[0,1]}|W_t^n - W_t^{n-1}| > \varepsilon_n\right) \le \exp(n\log 2 + 1/2 - 2^{(n+1)/2}\varepsilon_n).$$

If we set $\varepsilon_n = n^{-2}$, then evidently

$$\sum_{n=1}^{\infty}\mathbb{P}\left(\sup_{t\in[0,1]}|W_t^n - W_t^{n-1}| > \frac{1}{n^2}\right) < \infty$$

(this is why direct application of the Chebyshev inequality would have been too crude—we would not have been able to obtain this conclusion!) But by the Borel-Cantelli lemma, we find that this implies $\mathbb{P}(\sup_{t\in[0,1]}|W_t^n - W_t^{n-1}| > n^{-2}\text{ i.o.}) = 0$, so we have

$$\sup_{t\in[0,1]}|W_t^n - W_t^{n-1}| \le \frac{1}{n^2} \quad \text{for all } n \text{ sufficiently large} \quad \text{a.s.}$$

In particular, this implies that

$$\sum_{n=1}^{\infty}\sup_{t\in[0,1]}|W_t^n - W_t^{n-1}| < \infty \quad \text{a.s.}$$

But then we have a.s. uniform convergence by lemma 3.2.3, and so $W_t^n \to W_t$ as $n \to \infty$ a.s., where the process W_t has a.s. continuous sample paths. Nothing changes if we set the sample paths to zero in a null set which contains all discontinuous paths of W_t; this is an indistinguishable change, and now W_t has continuous sample paths everywhere. It remains to show that W_t has the correct finite dimensional distributions, and to extend to $t \in [0,\infty[$.

To verify the finite dimensional distributions, it suffices to show that for any $t > s > r$, the increment $W_t - W_s$ is independent of W_r, and that these are Gaussian random variables with mean zero and variance $t-s$ and r, respectively (why do we not need to check explicitly higher dimensional distributions?) The simplest way to check this is using characteristic functions: a well known result, e.g., [Wil91, section 16.6,7], states that it is sufficient to show that

$$\mathbb{E}(e^{i\alpha W_r + i\beta(W_t - W_s)}) = e^{-\alpha^2 r/2 - \beta^2(t-s)/2}.$$

But note that by construction, this holds for any t, s, r which are dyadic rationals (i.e., of the form $k2^{-n}$ for some k and n). For arbitrary t, s, r, choose sequences $r_n \nearrow r$, $s_n \searrow s$, and $t_n \searrow t$ of dyadic rationals. Then $W_{t_n} - W_{s_n}$ is independent of W_{r_n} for any n, and in particular we can calculate explicitly, using dominated convergence and continuity of W_t,

$$\begin{aligned}\mathbb{E}(e^{i\alpha W_r + i\beta(W_t - W_s)}) &= \lim_{n\to\infty}\mathbb{E}(e^{i\alpha W_{r_n} + i\beta(W_{t_n} - W_{s_n})}) \\ &= \lim_{n\to\infty} e^{-\alpha^2 r_n/2 - \beta^2(t_n - s_n)/2} = e^{-\alpha^2 r/2 - \beta^2(t-s)/2}.\end{aligned}$$

Hence W_t has the correct finite dimensional distributions for all $t \in [0,1]$. The extension to $t \in [0,\infty[$ was already done in lemma 3.2.1, and we finally have our Wiener process. □

Now that we have done the hard work of constructing a Wiener process, it is not difficult to prove proposition 3.1.2. Recall the statement of this result:

Proposition 3.1.2. *Suppose that we have constructed some stochastic process x_t whose finite dimensional distributions are those of lemma 3.1.1. Then there exists a modification $\tilde{x}_t$ of x_t such that $t \mapsto \tilde{x}_t$ is continuous.*

How would we go about proving this? Suppose that we have constructed a Wiener process W_t through theorem 3.2.5. It is an easy exercise to show that we can reproduce the random variables $\xi_{n,k}$ from W_t as follows:

$$\xi_{0,1} = W_1, \qquad \xi_{n,k} = 2^{(n+1)/2}(W_{k2^{-n}} - \tfrac{1}{2}W_{(k-1)2^{-n}} - \tfrac{1}{2}W_{(k+1)2^{-n}}), \quad n \geq 1.$$

But the law of any finite number of $\xi_{n,k}$ is only determined by the finite dimensional distributions of W_t, so for any process x_t with the same finite dimensional distributions it must also be the case that

$$\chi_{0,1} = x_1, \qquad \chi_{n,k} = 2^{(n+1)/2}(x_{k2^{-n}} - \tfrac{1}{2}x_{(k-1)2^{-n}} - \tfrac{1}{2}x_{(k+1)2^{-n}}), \quad n \geq 1,$$

are i.i.d. Gaussian random variables with mean zero and unit variance (recall that a sequence of random variables is independent if any finite subcollection is independent, so this notion only depends on the finite dimensional distributions). But then

$$\tilde{x}_t = \sum_{n=0}^{\infty} \sum_{k=1,3,\ldots,2^n-1} \chi_{n,k} S_{n,k}(t)$$

has a.s. continuous paths—this follows from the proof of theorem 3.2.5—and $x_t = \tilde{x}_t$ for all dyadic rational times t by construction. We again set the discontinuous paths to zero, and the only thing that remains to be shown is that $\tilde{x}_t$ is a modification of x_t.

Proof of proposition 3.1.2. We need to show that $\tilde{x}_t = x_t$ a.s. for fixed $t \in [0,1]$ (it suffices to restrict to $[0,1]$, as we can repeat the procedure for every interval $[n, n+1]$ separately). As with unit probability $\tilde{x}_t = x_t$ for all dyadic rational t and $\tilde{x}_t$ has continuous sample paths, we find $\tilde{x}_t = \lim_n \tilde{x}_{t_n} = \lim_n x_{t_n}$ a.s. for any sequence of dyadic rational times $t_n \nearrow t$. But

$$\begin{aligned}\mathbb{P}(|x_t - \tilde{x}_t| > \varepsilon) &\leq \varepsilon^{-2}\mathbb{E}((x_t - \tilde{x}_t)^2) = \varepsilon^{-2}\mathbb{E}(\liminf (x_t - x_{t_n})^2) \\ &\leq \varepsilon^{-2} \liminf \mathbb{E}((x_t - x_{t_n})^2) = \varepsilon^{-2} \liminf (t - t_n) = 0 \qquad \text{for any } \varepsilon > 0,\end{aligned}$$

where we have used Chebyshev's inequality and Fatou's lemma. Thus $x_t = \tilde{x}_t$ a.s. □

3.3 White noise

In the Introduction, we argued that the notion of white noise, as it is colloquially introduced in the science and engineering literature, can heuristically be thought of as the time derivative of the Wiener process. As was already mentioned, the nonexistence of white noise as a stochastic process will never be a problem, and we will happily consider noisy observations in their integrated form in order to avoid mathematical unpleasantness. Let us nonetheless take a moment now to look at white noise a little more closely. As we are already on the topic of the Wiener process, we should briefly investigate further the connection between the Wiener process and white noise.

In science and engineering, white noise is generally defined as follows: it is a Gaussian "stochastic process" ξ_t with zero mean and covariance $\mathbb{E}(\xi_s \xi_t) = \delta(t-s)$, where $\delta(\cdot)$ is Dirac's delta "function". The latter, however, is not actually a function;

it is a so-called *distribution*, or *generalized function*.[3] Let us briefly recall how this works. The delta function is defined by the relation

$$\int f(s)\,\delta(s)\,ds = f(0),$$

where f is an element in a suitable space of *test functions*. The simplest space of test functions is the space C_0^∞ of smooth functions of compact support. The mathematical object $\delta(\cdot)$ should then be seen not as a function, but as a linear functional on the space C_0^∞: it is a linear map which associates to every test function a number, in this case $\delta : f \mapsto f(0)$. The integral expression above is just suggestive notation[4] for δ evaluated at f. The philosophy behind such a concept is that no physical measurement can ever be infinitely sharp, even if the object which we are measuring is (which is itself an idealization); hence we only need to make sense of measurements that are smeared out in time by a suitable test function, and a generalized function is simply an object that associates to every such measurement the corresponding outcome.

Let us return to white noise. Clearly ξ_t is not a stochastic process, as its covariance is not a function. However, we could think of ξ_t as an object whose sample paths are themselves generalized functions. To make sense of this, we have to define the properties of white noise when integrated against a test function. So let us integrate the defining properties of white noise against test functions: $\mathbb{E}(\xi(f)) = 0$ and

$$\begin{aligned}\mathbb{E}(\xi(f)\xi(g)) &\equiv \mathbb{E}\left(\int_{\mathbb{R}_+} f(s)\,\xi_s\,ds \quad \int_{\mathbb{R}_+} g(t)\,\xi_t\,dt\right)\\ &= \int_{\mathbb{R}_+\times\mathbb{R}_+} f(s)\,g(t)\,\delta(t-s)\,ds\,dt = \int_{\mathbb{R}_+} f(t)\,g(t)\,dt \equiv \langle f, g\rangle.\end{aligned}$$

Moreover, the fact that ξ_t is a Gaussian "process" implies that $\xi(f)$ should be a Gaussian random variable for any test function f. So we can now define white noise as a *generalized* stochastic process: it is a random linear functional ξ on C_0^∞ such that $\xi(f)$ is Gaussian, $\mathbb{E}(\xi(f)) = 0$ and $\mathbb{E}(\xi(f)\xi(g)) = \langle f, g\rangle$ for every $f, g \in C_0^\infty$.

What is the relation to the Wiener process? The point of this section is to show that given a Wiener process W_t, the *stochastic integral*

$$\xi(f) = \int_0^\infty f(t)\,dW_t, \qquad f \in C_0^\infty,$$

satisfies the definition of white noise as a generalized stochastic process. This justifies to a large extent the intuition that stochastic integrals can be interpreted as integrals over white noise. It also justifies the idea of using the integrated observations

$$Y_t = \int_0^t a_s\,ds + W_t,$$

[3] We will prefer the name generalized function. The word distribution is often used in probability theory to denote the law of a random variable, not in the generalized function sense of L. Schwartz!

[4] The notation suggests that we can approximate $\delta(\cdot)$ by a sequence of actual functions $d_n(\cdot)$, such that the true (Riemann) integral $\int f(s)\,d_n(s)\,ds$ converges to $f(0)$ as $n \to \infty$ for every test function $f \in C_0^\infty$. This is indeed the case (think of a sequence of increasingly narrow normalized Gaussians).

rather than the engineering-style observations process $y_t = a_t + \xi_t$, as a model for a signal a_t corrupted by white noise; given Y_t, we could always reproduce the effect of a generalized function-style unsharp (smeared) measurement of y_t by calculating

$$y(f) = \int_0^\infty f(t)\, dY_t = \int_0^\infty f(t)\, a_t\, dt + \int_0^\infty f(t)\, dW_t.$$

The nice thing about stochastic integrals, however, is that they completely dispose of the need to work with generalized functions; the former live entirely within the domain of ordinary stochastic processes. As long as we are willing to accept that we sometimes have to work with integrated observations, rather than using white noise directly, what we gain is an extremely rich theory with very well developed analytical techniques (*stochastic calculus*). At the end of the day, you can still interpret these processes in white noise style (by smearing against a test function), without being constrained along the way by the many restrictions of the theory of generalized functions. Though white noise theory has its advocates—it is a matter of taste—it is fair to say that stochastic integrals have turned out to be by far the most fruitful and widely applicable. As such, you will not see another generalized function in this course.

To wrap up this section, it remains to show that the stochastic integral satisfies the properties of white noise. We have not yet introduced the stochastic integral, however; we previously broke off in desperation when we concluded that the infinite variation property of the Wiener process precludes the use of the Stieltjes integral for this purpose. Nonetheless we can rescue the Stieltjes integral for the purpose of this section, so that we can postpone the definition of a real stochastic integral until the next chapter. The reason that we do not get into trouble is that we only wish to integrate test functions in C_0^∞—as these functions are necessarily of finite variation (why?), we can define the stochastic integral through integration by parts. How does this work? Note that we can write for any partition π of $[0, T]$

$$\sum_{t_i \in \pi} f(t_i)\, (W_{t_{i+1}} - W_{t_i}) = f(T)\, W_T - \sum_{t_i \in \pi} W_{t_{i+1}}\, (f(t_{i+1}) - f(t_i)),$$

which is simply a rearrangement of the terms in the summation. But the sum on the right hand side limits to a Stieltjes integral: after all, f is a test function of finite variation, while W_t has continuous sample paths. So we can simply define

$$\int_0^T f(s)\, dW_s = f(T)\, W_T - \int_0^T W_s\, df(s),$$

where the integral on the right should be interpreted as a Stieltjes integral. In fact, as f is smooth, we can even write

$$\int_0^T f(s)\, dW_s = f(T)\, W_T - \int_0^T W_s\, \frac{df(s)}{ds}\, ds,$$

where we have used a well known property of the Stieltjes integral with respect to a continuously differentiable function. Our goal is to show that this functional has the properties of a white noise functional. Note that as f has compact support, we can simply define the integral over $[0, \infty[$ by choosing T to be sufficiently large.

Lemma 3.3.1. *The stochastic integral of $f \in C_0^\infty$ with respect to the Wiener process W_t (as defined through integration by parts) is a white noise functional.*

Proof. The integral is an a.s. limit (and hence a limit in distribution) of Gaussian random variables, so it must be itself a Gaussian random variable. It also has zero expectation: the only difficulty here is the exchange of the expectation and the Stieltjes integral, which is however immediately justified by Fubini's theorem. It remains to demonstrate the covariance identity. To this end, choose T to be sufficiently large so that the supports of both f and g are contained in $[0, T]$. Hence we can write (using $f(T) = g(T) = 0$)

$$\mathbb{E}(\xi(f)\xi(g)) = \mathbb{E}\left(\int_0^T W_s \frac{df(s)}{ds}\, ds \quad \int_0^T W_t \frac{dg(t)}{dt}\, dt\right).$$

Using Fubini's theorem and the elementary property $\mathbb{E}(W_s W_t) = s \wedge t$, we obtain

$$\mathbb{E}(\xi(f)\xi(g)) = \int_0^T \int_0^T (s \wedge t)\, df(s)\, dg(t).$$

The conclusion $\mathbb{E}(\xi(f)\xi(g)) = \langle f, g\rangle$ is an exercise in integration by parts. □

Unfortunately, this is about as far as the integration by parts trick will take us. In principle we could extend from test functions in C_0^∞ to test functions of finite variation, and we can even allow for random finite variation integrands. However, one of the main purposes of developing stochastic integrals is to have a stochastic calculus. Even if we naively try to apply the chain rule to calculate something like, e.g., W_t^2, we would still get integrals of the form $\int_0^T W_t\, dW_t$ which can never be given meaning through integration by parts. Hence we are really not going to be able to circumvent the limitations of the Stieltjes integral; ultimately, a different idea is called for.

3.4 Further reading

The Wiener process is both a classical topic is probability theory, and an active research topic to this day. It serves as the canonical example of a continuous time martingale and of a Markov process, has many fundamental symmetries, and its sample paths do not cease to fascinate. The sample path properties of the Wiener process are discussed, e.g., in the draft book by Mörters and Peres [MP06]. On the martingale side, the books by Karatzas and Shreve [KS91] and of Revuz and Yor [RY99] take the Wiener process as the starting point for the investigation of stochastic processes. The treatment in this chapter was largely inspired by Rogers and Williams [RW00a].

In the literature, the Wiener process is commonly constructed in three different ways. The first way is the most general-purpose, and is least specific to the properties of the Wiener process. There is a canonical method, the *Kolmogorov extention theorem*, using which a stochastic process can be constructed, on a suitable probability space, with specified finite dimensional distributions. The only requirement for this method is that the finite dimensional distributions satisfy certain natural consistency conditions. It is a priori unclear, however, whether such a process can be chosen to have continuous sample paths. Another result, *Kolmogorov's continuity theorem*, needs to be invoked to show that this can indeed be done. The latter gives conditions

on the finite dimensional distributions of a stochastic process under which a continuous modification is guaranteed to exist. See, e.g., Karatzas and Shreve [KS91].

The second way is through convergence of probability measures, as detailed in the classic text by Billingsley [Bil99]. This method begins with a sequence of random walks $w_t(n)$ with piecewise linear sample paths, each of which induces a measure μ_n on $C([0,1])$. We would like to show that the measures μ_n converge, in some sense, to a limiting measure μ on $C([0,1])$, the Wiener measure (under which the canonical process $\pi_t : x \mapsto x_t$ is a Wiener process). The appropriate notion of convergence is that of *weak convergence*, which means that $\mathbb{E}_{\mu_n}(f) \to \mathbb{E}_{\mu}(f)$ for every bounded function $f : C([0,1]) \to \mathbb{R}$ that is continuous in the topology of uniform convergence (this is precisely the correct notion of convergence in law for the stochastic processes $w_\cdot(n)$). To prove that the sequence of measures μ_n is actually weakly convergent, one needs the important notion of *tightness* which is beyond our scope.

Finally, you know the third (Lévy-Ciesielski) way—it is the one we have used. Incidentally, the method originally used by Wiener to construct his process is related, though not identical, to the approach which we have used. Rather than use Schauder functions with independent Gaussian coefficients, Wiener defined his process using a Fourier series with independent Gaussian coefficients.

In some sense we have not come full circle to the beginning of this chapter. Recall that we started with the idea that the Wiener process should be the limit of a sequence of random walks $x_t(N)$, where ξ_n were arbitrarily distributed i.i.d. random variables with zero mean and unit variance. In order to construct the Wiener process, however, we specialized this model considerably: we chose the ξ_n to be Gaussian, made the sample paths continuous, and constructed the walks very carefully. To make our story consistent we should show, now that we have constructed a Wiener process, that the more general random walks $x_t(N)$ still limit to a Wiener process. That this is indeed the case is the statement of *Donsker's invariance principle:*

Theorem 3.4.1 (Donsker). *$x_t(N)$ converges in law to a Wiener process W_t.*

As the random walks $x_t(N)$ do not have sample paths in $C([0,1])$, however, it is again unclear what we mean by convergence in law. In particular, we need to introduce a suitable topology on a larger space of functions which are allowed to have jumps, and show that for any bounded functional f that is continuous with respect to that topology we have $\mathbb{E}(f(x_\cdot(n))) \to \mathbb{E}(f(W_\cdot))$. The appropriate topology is the *Skorokhod topology*, which is described in detail in [Bil99]. If we were to choose our random walks to have piecewise linear sample paths, of course, then the compact uniform topology on $C([0,1])$ suffices and the result still holds.

As it will not be needed in the following, we do not prove Donsker's theorem here. There are two very different proofs of this theorem, for both of which you will find excellent discussions in the literature. The first proof uses the central limit theorem to prove weak convergence of the finite dimensional distributions (lemma 3.1.1), which is extended to weak convergence of the entire process using some analytic arguments; see [Bil99, theorems 7.5,8.2,14.1]. The second method uses the *Skorokhod embedding* to express a random walk in terms of a Wiener process evaluated a sequence of stopping times; see [Kal97, chapter 12]. The latter does not use the central limit theorem, and in fact the central limit theorem follows from this technique as a corollary.

CHAPTER 4

The Itô Integral

The stage is finally set for introducing the solution to our stochastic integration problem: the *Itô integral*, and, of equal importance, the associated *stochastic calculus* which allows us to manipulate such integrals.

4.1 What is wrong with the Stieltjes integral?

Before we define the Itô integral, let us take a little closer look at the Stieltjes integral. We stated (without proof) in section 3.1 that if the integrator of a Stieltjes integral is of infinite variation, then there is a continuous integrand for which the integral is not defined. It is difficult, however, to construct an explicit example. Instead, we will take a slightly different approach to the Stieltjes integral in this section, and show how it gets us into trouble. If we can figure out precisely what goes wrong, this should give an important hint as to what we need to do to resolve the integration problem.

Remark 4.1.1. This section attempts to explain why the Itô integral is defined the way it is. You do not need to read this section in order to understand the Itô integral; skip to the next section if you wish to get started right away!

The Stieltjes integral revisited

For sake of example, let f and g be continuous functions on the interval $[0, 1]$. How should we define the Stieltjes integral of f with respect to g? In section 3.1 we did this in "Riemann fashion" (i.e., as in the definition of the Riemann integral) by sampling the function f on an increasingly fine partition. Let us be a little more general, however, and define the Stieltjes integral in a manner closer to the definition of the Lebesgue integral, i.e., by defining the integral first for simple functions and then ex-

tending the definition by taking limits. To this end, let f_n be a simple function, i.e., one that is piecewise constant and jumps at a finite number of times t_i^n. Define

$$I(f_n) = \int_0^1 f_n(s)\, dg(s) = \sum_i f_n(t_i^n)\,(g(t_{i+1}^n) - g(t_i^n)),$$

where we have evaluated the integrand at t_i^n for concreteness (any point in the interval $[t_i^n, t_{i+1}^n]$ should work). Now choose the sequence of simple functions $\{f_n\}$ so that it converges uniformly to f, i.e., $\sup_{t\in[0,1]} |f_n(t) - f(t)| \to 0$ as $n \to \infty$. Then

$$I(f) = \int_0^1 f(s)\, dg(s) = \lim_{n\to\infty} I(f_n) = \lim_{n\to\infty} \int_0^1 f_n(s)\, dg(s).$$

Does this definition make sense? We should verify two things. First, we need to show that the limit exists. Second, we need to show that the limit is independent of how we choose our simple approximations f_n.

Lemma 4.1.2. *Suppose that g has finite variation $\mathrm{TV}(g,0,1) < \infty$ and that the sequence of simple functions f_n converges to f uniformly. Then the sequence $I(f_n)$ converges, and its limit does not depend on the choice of the approximations f_n.*

Proof. First, choose some fixed m, n, and let t_i be the sequence of times that includes the jump times t_i^m and t_i^n of both f_m and f_n, respectively. Then clearly

$$I(f_n) - I(f_m) = \sum_i (f_n(t_i) - f_m(t_i))\,(g(t_{i+1}) - g(t_i)),$$

so that in particular

$$|I(f_n) - I(f_m)| \le \sum_i |f_n(t_i) - f_m(t_i)|\,|g(t_{i+1}) - g(t_i)| \le \sup_t |f_n(t) - f_m(t)|\,\mathrm{TV}(g,0,1).$$

As $\sup_{t\in[0,1]} |f_n(t) - f(t)| \to 0$, we find that $|I(f_n) - I(f_m)| \to 0$ as $m, n \to \infty$. Evidently $I(f_n)$ is a Cauchy sequence in $\mathbb{R}$, and hence converges. It remains to show that the limit does not depend on the choice of approximation. To this end, let h_n be another sequence of simple functions that converges uniformly to f. Then, by the same argument, $|I(h_n) - I(f_n)| \to 0$ as $n \to \infty$, which establishes the claim. □

Remark 4.1.3. The Riemann-type definition of section 3.1 and the current definition coincide: the former corresponds to a particular choice of simple approximations.

We have seen nothing that we do not already know. The question is, what happens when $\mathrm{TV}(g,0,1) = \infty$, e.g., if g is a typical sample path of the Wiener process? The main point is that in this case, the previous lemma fails miserably. Let us show this.

Lemma 4.1.4. *Suppose that g has infinite variation $\mathrm{TV}(g,0,1) = \infty$. Then there exist simple functions f_n which converge to f uniformly, such that $I(f_n)$ diverges.*

Proof. It suffices consider $f = 0$. After all, suppose that f_n converges uniformly to f such that $I(f_n)$ converges; if h_n is a sequence of simple functions that converges to zero uniformly, such that $I(h_n)$ diverges, then $f_n + h_n$ also converges uniformly to f but $I(f_n + h_n)$ diverges.

As g has infinite variation, there exists a sequence of partitions π_n of $[0, 1]$ such that

$$\sum_{t_i \in \pi_n} |g(t_{i+1}) - g(t_i)| \to \infty \quad \text{as } n \to \infty.$$

Define h_n to be a simple function such that $h_n(t_i) = \text{sign}(g(t_{i+1}) - g(t_i))$ for all $t_i \in \pi_n$. Evidently $I(h_n) \to \infty$, but certainly h_n does not converge uniformly. However, define the sequence of simple functions $f_n = (I(h_n))^{-1/2} h_n$. Then $I(f_n) = (I(h_n))^{1/2} \to \infty$ as well, but clearly $\sup_t |f_n(t)| = (I(h_n))^{-1/2} \to 0$, so f_n converges uniformly to zero. □

Apparently we cannot define the Stieltjes integral unambiguously if g has infinite variation: *the integral will depend on the choice of the approximating sequence!* In fact, beside making the integral diverge, we can make the integral converge to whatever we like if we use $f_n \propto (I(h_n))^{-1} h_n$ in the previous proof.

Maybe things are not as bad as they seem

When g is a Wiener process, however, the negative conclusion of the previous lemma is not so threatening—at least not if we choose h_n to be non-random. To see this, let h_n be any (deterministic) simple function that takes the values ± 1 and switches at the jump times t_i. Then $h_n(t_i) = \text{sign}(W_{t_{i+1}} - W_{t_i})$ only with probability one half: after all, $W_{t_{i+1}} - W_{t_i}$ is Gaussian with mean zero and so has either sign with equal probability. Moreover, you can easily see that

$$\mathbb{P}(h_n(t_i) = \text{sign}(W_{t_{i+1}} - W_{t_i}) \ \forall\, i) = 2^{-\#\Delta h_n},$$

where $\#\Delta h_n$ is the number of jumps of h_n (this follows from the fact that W_t has independent increments). In fact, it follows from the Borel-Cantelli lemma in this case that $\mathbb{P}(h_n(t_i) = \text{sign}(W_{t_{i+1}} - W_{t_i}) \ \forall\, i \text{ i.o.}) = 0$, regardless of how we choose the sequence h_n (provided $\#\Delta h_n$ increases when n increases). Though this does not prove anything in itself, it suggests that things might not be as bad as they seem: lemma 4.1.4 shows that there are certain sample paths of W_t for which the integral of f_n diverges, but it seems quite likely that the set of all such sample paths is always of probability zero! In that case (and this is indeed the case[1]) we are just fine—we only care about defining stochastic integrals with probability one.

Unfortunately, we are not so much interested in integrating deterministic integrands against a Wiener process: we would like to be able to integrate random processes. In this case we are in trouble again, as we can apply lemma 4.1.4 for every sample path separately[2] to obtain a uniformly convergent sequence of stochastic processes f_n whose integral with respect to W_t diverges.

[1] From the discussion below on the Wiener integral, it follows that the integral of f_n converges to zero in $\mathcal{L}^2$. Hence the integral can certainly not diverge with nonzero probability.

[2] For every $\omega \in \Omega$ separately, set $g(t) = W_t(\omega)$ apply lemma 4.1.4 to obtain $f_n(t, \omega)$. There is a technical issue here (is the stochastic process f_n measurable?), but this can be resolved with some care.

A way out

The proof of lemma 4.1.4 suggests a way out. The key to the proof of lemma 4.1.4 was that we could construct an offensive sequence f_n by "looking into the future": f_n is constructed so that its sign matches the sign of the future increment of g. By doing this, we can express the total variation of g as a limit of simple integrals, so that the integral diverges whenever g has infinite variation.

This cunning trick is foiled, however, if we make g a Wiener process but keep f_n non-random: in that case we can never look into the future, because $f_n(t_i)$, being non-random, cannot contain any information on the sign of $W_{t_{i+1}} - W_{t_i}$. Even if f_n were allowed to be random, however, this would still be the case if we require $f_n(t_i)$ to be *independent* of $W_{t_{i+1}} - W_{t_i}$! Fortunately enough, there is a rich and important class of stochastic processes with precisely this property.

Key idea 1. Let W_t be an $\mathcal{F}_t$-Wiener process. Then we will only define stochastic integrals with respect to W_t of stochastic processes which are *$\mathcal{F}_t$-adapted*.

This key idea puts an end to the threat posed by lemma 4.1.4. But it is still not clear how we should proceed to actually define the stochastic integral: after all, lemma 4.1.2 does not (and can not, by lemma 4.1.4) hold water in the infinite variation setting.

Things look nicer in mean square

The second key idea comes from the proof of lemma 3.1.11. There we saw that even though the finite variation of the Wiener process is a.s. infinite, the *quadradic variation* is finite: for any refining sequence of partitions π_n of the interval $[0, 1]$, we have

$$\sum_{t_i \in \pi_n} (W_{t_{i+1}} - W_{t_i})^2 \to 1 \quad \text{in probability.}$$

This suggests that we might be able to repeat the proof of lemma 4.1.2 using convergence *in* $\mathcal{L}^2$ rather than a.s. convergence, exploiting the finiteness of the *quadratic* variation rather than the total variation. In particular, we can try to prove that the sequence of simple integrals $I(f_n)$ is a Cauchy sequence in $\mathcal{L}^2$, rather than proving that $I(f_n)(\omega)$ is a Cauchy sequence in $\mathbb{R}$ for almost every ω. This indeed turns out to work, provided that we stick to adapted integrands as above.

Key idea 2. As the *quadratic* variation of the Wiener process is finite, we should define the stochastic integrals as limits *in* $\mathcal{L}^2$.

Let us show that this actually works in the special case of non-random integrands. If f_n is a non-random simple function on $[0, 1]$, then, as usual,

$$I(f_n) = \int_0^1 f_n(s)\, dW_s = \sum_i f_n(t_i^n)\, (W_{t_{i+1}^n} - W_{t_i^n}),$$

where t_i^n are the jump times of f_n. Using the independence of the increments of the Wiener process, we obtain immediately

$$\mathbb{E}((I(f_n))^2) = \sum_i (f_n(t_i^n))^2\, (t_{i+1}^n - t_i^n) = \int_0^1 (f_n(s))^2\, ds,$$

and in particular we find that

$$\mathbb{E}((I(f_n) - I(f_m))^2) = \int_0^1 (f_n(s) - f_m(s))^2 \, ds.$$

Now suppose that the functions f_n converge to some function f in $\mathcal{L}^2([0,1])$, i.e.,

$$\int_0^1 (f_n(s) - f(s))^2 \, ds \to 0 \quad \text{as} \quad n \to \infty.$$

Then $\mathbb{E}((I(f_n) - I(f_m))^2) \to 0$ as $m, n \to \infty$, i.e., the sequence $I(f_n)$ is Cauchy in $\mathcal{L}^2$. Hence $I(f_n)$ converges in $\mathcal{L}^2$ to some random variable $I(f)$. A simple argument shows that the integral thus defined does not depend on the choice of approximations, and thus we have defined a bona fide integral called the *Wiener integral*.

Remark 4.1.5. The Wiener integral is only defined for non-random integrands. Recall that we previously encountered another such integral: see lemma 3.3.1. You can easily verify that the latter is a special case of the Wiener integral, obtained by restricting the class of integrands to be the smooth functions.

Remark 4.1.6. The finiteness of the quadratic variation is what makes this procedure work. However, the quadratic variation is a little hidden in the above discussion, and indeed we will not often see it appear explicitly in this chapter. It is possible, however, to extend these ideas to the case where the integrator is an arbitrary martingale (rather than a Wiener process). In that case the quadratic variation shows up very explicitly in the construction. As this is an introductory course, we will not pursue this direction; suggestions for further reading can be found in section 4.7.

You might worry that the limit in $\mathcal{L}^2$ clashes with the interpretation of the stochastic integral in real-world applications. For example, when we define stochastic differential equations in the next chapter, we will think of these as idealizations of ordinary differential equations driven by rapidly fluctuating noise. In every run of an experiment, however, only one sample path of the noise occurs, and the system represented by the differential equation is really driven by that particular sample path in that run of the experiment. On the other hand, the limit in $\mathcal{L}^2$ suggests that we cannot think of every sample path of the noise individually—ostensibly we can only think of all of them together in some average sense, as though the sample paths of the noise that do not occur in the current realization can somehow influence what is currently happening. In other words, it is not clear that we can actually compute the value of the integral $I(f)$, given only the sample path $W_t(\omega)$ that is realized in the current run of the experiment, without involving the other sample paths $W_t(\omega')$ in the matter.

This is not really an issue, however; in fact, it is mostly a matter of definition. If we wish, we can still define the Wiener integral as an a.s. limit rather than a limit in $\mathcal{L}^2$: all we have to do is require that our sequence of simple functions f_n obeys

$$\sum_{n=1}^{\infty} \int_0^1 (f_n(s) - f(s))^2 \, ds < \infty,$$

i.e., that it converges to f_n fast enough. (The arguments that lead to a.s. convergence should be very familiar to you by now!) This way we really obtain the integral for every sample path of the noise separately, and the conceptual issues are resolved. Exactly the same holds for the Itô integral, to be defined in the next section. Note that this does not change the nature of the stochastic integrals—they are still fundamentally limits in $\mathcal{L}^2$, as can be seen by the requirement that $f_n \to f$ in $\mathcal{L}^2([0,1])$. The point is merely that this does not preclude the *pathwise* computation of the integral (as is most natural from a conceptual point of view). In the following we will thus not worry about this issue, and define integrals as limits in $\mathcal{L}^2$ without further comment.

What can happen if we do not take $f_n \to f$ fast enough? You can indeed construct examples where $f_n \to 0$ in $\mathcal{L}^2([0,1])$, but the Wiener integral $I(f_n)$ does not converge a.s. (of course, it does converge to zero in $\mathcal{L}^2$). Consider, for example, the simple functions $f_n = H_{n,1}/\alpha_n$, where $H_{n,1}$ is the Haar wavelet constructed in theorem 3.2.5, and $\alpha_n > 0$ is chosen such that $\mathbb{P}(|\xi| > \alpha_n) = n^{-1}$ where ξ is a Gaussian random variable with zero mean and unit variance. Then $\alpha_n \to \infty$ as $n \to \infty$, so $f_n \to 0$ in $\mathcal{L}^2([0,1])$. On the other hand, $I(H_{n,k})$ are i.i.d. Gaussian random variables with zero mean and unit variance (see the discussion after the proof of theorem 3.2.5), so some set manipulation gives (how?)

$$\mathbb{P}(|I(f_n)| > 1 \text{ i.o.}) = 1 - \lim_{m\to\infty} \prod_{n\geq m} (1 - \mathbb{P}(|I(H_{n,1})| > \alpha_n)) \leq 1 - \lim_{m\to\infty} e^{-\sum_{n\geq m} n^{-1}} = 1$$

where we have used the estimate $1 - x \leq e^{-x}$. Hence $I(f_n)$ certainly cannot converge a.s.

The behavior of the integral in this case is equivalent to that of example 1.5.6, i.e., there is an occasional excursion of $I(f_n)$ away from zero which becomes increasingly rare as n gets large (otherwise $I(f_n)$ would not converge in $\mathcal{L}^2$). This need not pose any conceptual problem; you may simply consider it an artefact of a poor choice of approximating sequence.

4.2 The Itô integral

Throughout this section, let us fix a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\in[0,\infty[}, \mathbb{P})$ and an $\mathcal{F}_t$-Wiener process W_t. We are going to define stochastic integrals with respect to W_t. The clues that we obtained from the previous section are

1. we should consider only $\mathcal{F}_t$-adapted integrands; and
2. we should try to take limits in $\mathcal{L}^2$.

We will construct the integral in two passes. First, we will develop a minimal construction of the integral which emphasizes the basic ideas. In the second pass, we will add some bells and whistles that make the Itô integral the truly powerful tool that it is.

A bare-bones construction

Let $\{X_t^n\}_{t\in[0,T]}$ be a simple, square-integrable, $\mathcal{F}_t$-adapted stochastic process. What does this mean? In principle we could allow every sample path $X_t^n(\omega)$ to have its own jump times $t_i(\omega)$, but for our purposes it will suffice to assume that the jump times

t_i are non-random.[3] So we assume that X_t^n is a constant, $\mathcal{F}_{t_i}$-measurable random variable in $\mathcal{L}^2$ (i.e., square-integrable) for $t_i \leq t < t_{i+1}$, where t_i, $i = 0, \ldots, N+1$ is a finite set of non-random jump times (with our usual convention that $t_0 = 0$ and $t_{N+1} = T$). For such simple integrands, we define the stochastic integral

$$I(X_\cdot^n) = \int_0^T X_t^n \, dW_t = \sum_{i=0}^{N} X_{t_i}^n \, (W_{t_{i+1}} - W_{t_i}),$$

and our goal is to extend this definition to a more general class of integrands by taking limits. To this end, we will need the following *Itô isometry:*

$$\mathbb{E}\left[\left(\int_0^T X_t^n \, dW_t\right)^2\right] = \sum_{i=0}^{N} \mathbb{E}((X_{t_i}^n)^2)\,(t_{i+1} - t_i) = \mathbb{E}\left[\int_0^T (X_t^n)^2 \, dt\right].$$

Note that it is crucial that X_t^n is $\mathcal{F}_t$-adapted: because of this $X_{t_i}^n$ is $\mathcal{F}_{t_i}$-measurable, so is independent of $W_{t_{i+1}} - W_{t_i}$, and this is absolutely necessary for the Itô isometry to hold! The requirement that $X_t^n \in \mathcal{L}^2$ is also necessary at this point, as otherwise $\mathbb{E}((X_t^n)^2)$ would not be finite and we would run into trouble.

Let us look a little more closely at the various objects in the expressions above. The Itô integral $I(X_\cdot^n)$ is a random variable in $\mathcal{L}^2(\mathbb{P})$. It is fruitful to think of the stochastic process X_t^n as a measurable map $X_\cdot^n : [0,T] \times \Omega \to \mathbb{R}$. If we consider the product measure $\mu_T \times \mathbb{P}$ on $[0,T] \times \Omega$, where μ_T is the Lebesgue measure on $[0,T]$ (i.e., T times the uniform probability measure), then the right-hand side of the Itô isometry is precisely $\mathbb{E}_{\mu_T \times \mathbb{P}}((X_\cdot^n)^2)$. In particular, the Itô isometry reads

$$\|I(X_\cdot^n)\|_{2,\mathbb{P}} = \|X_\cdot^n\|_{2,\mu_T \times \mathbb{P}},$$

where $\|\cdot\|_{2,\mathbb{P}}$ is the $\mathcal{L}^2$-norm on Ω and $\|\cdot\|_{2,\mu_T \times \mathbb{P}}$ is the $\mathcal{L}^2$-norm on $[0,T] \times \Omega$. This is precisely the reason for the name *isometry*—the mapping $I : \mathcal{L}^2(\mu_T \times \mathbb{P}) \to \mathcal{L}^2(\mathbb{P})$ preserves the $\mathcal{L}^2$-distance (i.e., $\|I(X_\cdot^n) - I(Y_\cdot^n)\|_{2,\mathbb{P}} = \|X_\cdot^n - Y_\cdot^n\|_{2,\mu_T \times \mathbb{P}}$), *at least when applied to $\mathcal{F}_t$-adapted simple integrands*. This fact can now be used to extend the definition of the Itô integral to a larger class of integrands in $\mathcal{L}^2(\mu_T \times \mathbb{P})$.

Lemma 4.2.1. *Let $X_\cdot \in \mathcal{L}^2(\mu_T \times \mathbb{P})$, and suppose there exists a sequence of $\mathcal{F}_t$-adapted simple processes $X_\cdot^n \in \mathcal{L}^2(\mu_T \times \mathbb{P})$ such that*

$$\|X_\cdot^n - X_\cdot\|_{2,\mu_T \times \mathbb{P}}^2 = \mathbb{E}\left[\int_0^T (X_t^n - X_t)^2 \, dt\right] \xrightarrow{n \to \infty} 0.$$

Then $I(X_\cdot)$ can be defined as the limit in $\mathcal{L}^2(\mathbb{P})$ of the simple integrals $I(X_\cdot^n)$, and the definition does not depend on the choice of simple approximations $X_\cdot^n$.

[3] In a more general theory, where we can integrate against arbitrary martingales instead of the Wiener process, the sample paths of the integrator could have jumps. In that case, it can become necessary to make the jump times t_i of the simple integrands random, and we have to be more careful about whether the integrand and integrator are left- or right-continuous at the jumps (see [Pro04]). This is not an issue for us.

Proof. As $\|X_\cdot^n - X_\cdot\|_{2,\mu_T\times\mathbb{P}} \to 0$ as $n \to \infty$, we find that

$$\|X_\cdot^n - X_\cdot^m\|_{2,\mu_T\times\mathbb{P}} \le \|X_\cdot^n - X_\cdot\|_{2,\mu_T\times\mathbb{P}} + \|X_\cdot^m - X_\cdot\|_{2,\mu_T\times\mathbb{P}} \to 0 \quad \text{as} \quad m,n\to\infty.$$

But all X^n are adapted, so the Itô isometry gives $\|I(X_\cdot^n) - I(X_\cdot^m)\|_{2,\mathbb{P}} \to 0$ as $m,n\to\infty$. Hence $I(X_\cdot^n)$ is a Cauchy sequence in $\mathcal{L}^2(\mathbb{P})$, and we denote by $I(X_\cdot)$ its limit in $\mathcal{L}^2(\mathbb{P})$. To prove uniqueness, let $Y_\cdot^n \in \mathcal{L}^2(\mu_T\times\mathbb{P})$ be another sequence of $\mathcal{F}_t$-adapted simple processes such that $\|Y_\cdot^n - X_\cdot\|_{2,\mu_T\times\mathbb{P}} \to 0$, and denote by $I(Y_\cdot)$ the limit in $\mathcal{L}^2(\mathbb{P})$ of $I(Y_\cdot^n)$. Then

$$\|I(Y_\cdot) - I(X_\cdot)\|_{2,\mathbb{P}} \le \|I(Y_\cdot) - I(Y_\cdot^n)\|_{2,\mathbb{P}} + \|I(Y_\cdot^n) - I(X_\cdot^n)\|_{2,\mathbb{P}} + \|I(X_\cdot^n) - I(X_\cdot)\|_{2,\mathbb{P}},$$

where the first and last terms on the right converge to zero by definition, while the fact that the second term converges to zero follows easily from the Itô isometry. Hence $I(Y_\cdot) = I(X_\cdot)$ a.s., so the integral does not depend on the approximating sequence. □

The question thus becomes, which stochastic processes $X_\cdot \in \mathcal{L}^2(\mu_T\times\mathbb{P})$ can actually be approximated by simple, $\mathcal{F}_t$-adapted processes? It turns out that this is the case for *any* $\mathcal{F}_t$-adapted $X_\cdot \in \mathcal{L}^2(\mu_T\times\mathbb{P})$.

Lemma 4.2.2. *Let $X_\cdot \in \mathcal{L}^2(\mu_T\times\mathbb{P})$ be $\mathcal{F}_t$-adapted. Then there exists a sequence of $\mathcal{F}_t$-adapted simple processes $X_\cdot^n \in \mathcal{L}^2(\mu_T\times\mathbb{P})$ such that $\|X_\cdot^n - X_\cdot\|_{2,\mu_T\times\mathbb{P}} \to 0$.*

Proof. Suppose that $X_\cdot$ has continuous and bounded sample paths. Then the simple functions

$$X_t^n = X_{k2^{-n}T}, \qquad k2^{-n}T \le t < (k+1)2^{-n}T, \qquad k = 0,\dots,2^n-1,$$

converge to X_t for every sample path separately; after all, as $[0,T]$ is compact, the sample paths are uniformly continuous, so $\sup_{t\in[0,T]}|X_t^n - X_t| \le \sup_t \sup_{s\in[0,2^{-n}T]}|X_t - X_{t+s}| \to 0$ as $n\to\infty$. But as the sample paths are uniformly bounded, it follows that $X_\cdot^n \to X_\cdot$ in $\mathcal{L}^2(\mu_T\times\mathbb{P})$ by the dominated convergence theorem.

Now suppose that $X_\cdot$ is just bounded and progressively measurable. Define the process

$$X_t^\varepsilon = \frac{1}{\varepsilon}\int_{t-\varepsilon}^t X_{s\vee 0}\,ds.$$

Then $X_\cdot^\varepsilon$ has bounded, continuous sample paths, is $\mathcal{F}_t$-adapted (by the progressive measurability of $X_\cdot$), and $X_t^\varepsilon \to X_t$ as $\varepsilon\to 0$ for every sample path separately. For any $\varepsilon > 0$, we can thus approximate X_t^ε by simple adapted processes $X_t^{n,\varepsilon}$, so by dominated convergence

$$\lim_{\varepsilon\to 0}\lim_{n\to\infty}\mathbb{E}\left[\int_0^T (X_t^{n,\varepsilon} - X_t)^2\,dt\right] = 0.$$

As such, we can find a subsequence $\varepsilon_n \searrow 0$ such that $X_\cdot^{n,\varepsilon_n} \to X_\cdot$ in $\mathcal{L}^2(\mu_T\times\mathbb{P})$.

Next, suppose that $X_\cdot$ is just progressively measurable. Then the process $X_t I_{|X_t|\le M}$ is progressively measurable and bounded, and, moreover,

$$\lim_{M\to\infty}\mathbb{E}\left[\int_0^T (X_t I_{|X_t|\le M} - X_t)^2\,dt\right] = \lim_{M\to\infty}\mathbb{E}\left[\int_0^T (X_t)^2 I_{|X_t|>M}\,dt\right] = 0$$

by the dominated convergence theorem. As before, we can find a sequence of simple adapted processes $X_t^{n,M}$ that approximate $X_t I_{|X_t|\le M}$ as $n\to\infty$, and hence there is a subsequence $M_n \nearrow \infty$ such that $X_\cdot^{n,M_n} \to X_\cdot$ in $\mathcal{L}^2(\mu_T\times\mathbb{P})$, as desired.

The result can be exteded further even to the case that $X_\cdot$ is not progressively measurable. Such generality is not of much interest to us, however; see [KS91, lemma 3.2.4]. □

We have now constructed the Itô integral in its most basic form. To be completely explicit, let us state what we have learned as a definition.

Definition 4.2.3 (Elementary Itô integral). Let X_t be any $\mathcal{F}_t$-adapted process in $\mathcal{L}^2(\mu_T \times \mathbb{P})$. Then the Itô integral $I(X_\cdot)$, defined as the limit in $\mathcal{L}^2(\mathbb{P})$ of simple integrals $I(X^n_\cdot)$, exists and is unique (i.e., is independent of the choice of X^n_t).

Before we move on, let us calculate a simple example "by hand." (This example will become completely trivial once we have the Itô calculus!)

Example 4.2.4. We would like to calculate the integral of W_t with respect to itself. As W_t has continuous sample paths, we find that

$$\int_0^T W_t \, dW_t = \mathcal{L}^2\text{-}\lim_{n\to\infty} \sum_{k=0}^{2^n-1} W_{k2^{-n}T} \, (W_{(k+1)2^{-n}T} - W_{k2^{-n}T}).$$

But note that we can rearrange the sum as

$$2 \sum_{k=0}^{2^n-1} W_{k2^{-n}T} \, (W_{(k+1)2^{-n}T} - W_{k2^{-n}T}) = W_T^2 - \sum_{k=0}^{2^n-1} (W_{(k+1)2^{-n}T} - W_{k2^{-n}T})^2,$$

and the second term on the right converges in $\mathcal{L}^2$ to T (recall that this is precisely the quadratic variation of W_t). So we find that

$$W_T^2 = 2 \int_0^T W_t \, dW_t + T.$$

Certainly this would not be the case if the Itô integral were a Stieltjes integral; the ordinary chain rule suggests that $d(W_t)^2/dt = 2W_t \, dW_t/dt$!

The full-blown Itô integral

We have now constructed the Itô integral; what more is there to do? We have two issues to take care of in this section which are well worth the effort.

First, in the previous discussion we had fixed a terminal time T, and defined the integral over the interval $[0, T]$. However, we are usually interested in considering the Itô integral as a stochastic process: i.e., we would like to consider the process

$$t \mapsto \int_0^t X_s \, dW_s, \qquad t \in [0, \infty[.$$

We will show in this section that we can choose the Itô integral so that it has continuous sample paths, a task which is reminiscent of our efforts in defining the Wiener process. Continuity is usually included, implicitly or explicitly, in the definition of the Itô integral—we will always assume it throughout this course.

The second issue is that the class of processes which we can integrate using the elementary Itô integral—the $\mathcal{F}_t$-adapted processes in $\mathcal{L}^2(\mu_T \times \mathbb{P})$—is actually a little too restrictive. This may seem on the whiny side; surely this is a huge class of stochastic

processes? But here is the problem. We will shortly be setting up a stochastic calculus, which will allow us to express functions of Itô integrals as new Itô integrals. For example, given two Itô integrals $I(X_\cdot)$ and $I(Y_\cdot)$, we will obtain a rule which allows us to express the product $I(X_\cdot)I(Y_\cdot)$ as the sum of a single Itô integral $I(Z_\cdot)$ of some process $Z_\cdot$ and a time integral. Even if $X_\cdot$ and $Y_\cdot$ are in $\mathcal{L}^2(\mu_T \times \mathbb{P})$, however, this does not guarantee that the appropriate process $Z_\cdot$ will be in $\mathcal{L}^2(\mu_T \times \mathbb{P})$. By extending the class of integrable processes, we will make sure that the the product of two Itô integrals can always be expressed as another Itô integral. This ultimately makes the theory easier to use, as you do not have to think, every time you wish to manipulate an Itô integral, whether that particular manipulation is actually allowed.

We proceed as follows. We first prove that for adapted integrands in $\mathcal{L}^2(\mu_T \times \mathbb{P})$, we can define the Itô integral as a stochastic process on $[0, T]$ with continuous sample paths. Next, we define the Itô integral as a stochastic process on $[0, \infty[$ with continuous paths, by extending our previous construction through a simple process called *localization*. By modifying the localization trick just a little bit, we will subsequently be able to extend the Itô integral to a much larger class of integrands.

Continuous sample paths

How to define the Itô integral with continuous sample paths? We will try to apply our standard trick which served us so well in constructing the Wiener process. First we define the simple integrals so that they have continuous sample paths; then we prove that there exists a subsequence of the simple integrals that converges uniformly a.s. The limiting sample paths are then automatically continuous.

Let X_t^n be an $\mathcal{F}_t$-adapted simple process in $\mathcal{L}^2(\mu_T \times \mathbb{P})$ with jump times t_i. For any time $t \leq T$, we define the simple integral

$$I_t(X_\cdot^n) = \int_0^t X_s^n \, dW_s = \int_0^T X_s^n I_{s \leq t} \, dW_s = \sum_{i=0}^N X_{t_i}^n \, (W_{t_{i+1} \wedge t} - W_{t_i \wedge t}).$$

The stochastic process $I_t(X_\cdot^n)$ has continuous sample paths; this follows immediately from the fact that the Wiener process has continuous sample paths. The simple integral has another very important property, however—it is an $\mathcal{F}_t$*-martingale*.

Lemma 4.2.5. *$I_t(X_\cdot^n)$ is an $\mathcal{F}_t$-martingale.*

Proof. If $r \leq t_i < t$, then

$$\mathbb{E}(X_{t_i}^n \, (W_{t_{i+1} \wedge t} - W_{t_i \wedge t}) | \mathcal{F}_r) = \mathbb{E}(X_{t_i}^n \, \mathbb{E}(W_{t_{i+1} \wedge t} - W_{t_i \wedge t} | \mathcal{F}_{t_i}) | \mathcal{F}_r) = 0,$$

as $W_{t_{i+1} \wedge t} - W_{t_i \wedge t}$ is independent of $\mathcal{F}_{t_i}$. If $t_i < r < t_{i+1} \wedge t$, then

$$\mathbb{E}(X_{t_i}^n \, (W_{t_{i+1} \wedge t} - W_{t_i \wedge t}) | \mathcal{F}_r) = X_{t_i}^n \, \mathbb{E}(W_{t_{i+1} \wedge t} - W_{t_i \wedge t} | \mathcal{F}_r) = X_{t_i}^n \, (W_r - W_{t_i}),$$

whereas for $r \geq t_{i+1} \wedge t$ clearly $\mathbb{E}(X_{t_i}^n \, (W_{t_{i+1} \wedge t} - W_{t_i \wedge t}) | \mathcal{F}_r) = X_{t_i}^n \, (W_{t_{i+1} \wedge t} - W_{t_i \wedge t})$. Hence for any $r < t$, $\mathbb{E}(I_t(X_\cdot^n) | \mathcal{F}_r)$ can be calculated as

$$\sum_{i=0}^N \mathbb{E}(X_{t_i}^n \, (W_{t_{i+1} \wedge t} - W_{t_i \wedge t}) | \mathcal{F}_r) = \sum_{i=0}^N X_{t_i}^n \, (W_{t_{i+1} \wedge r} - W_{t_i \wedge r}) = I_r(X_\cdot^n),$$

which is the martingale property. Hence we are done. □

Remark 4.2.6. The fact the the simple integral is a martingale should not come as a surprise—the discrete time process $I_{t_i}(X^n_\cdot)$ is a martingale transform! We can interpret the Itô integral as a continuous time martingale transform, albeit for a very specific martingale: the Wiener process. (The more general theory, where you can integrate against any martingale, makes this interpretation even more convincing.)

The martingale property is extremely helpful in constructing continuous sample paths. To accomplish the latter, we will copy almost literally the argument used in the proof of theorem 3.2.5 to construct the Wiener process with continuous sample paths. That argument, however, relied on an estimate which, in the current context, corresponds to a bound on $\mathbb{P}(\sup_{t\in[0,T]}|I_t(X^n_\cdot)| > \varepsilon_n)$. The martingale property provides an ideal tool to obtain such bounds: we can simply copy the argument that led to the proof of the supermartingale inequality, lemma 2.3.21.

Lemma 4.2.7. *Let X_t be an $\mathcal{F}_t$-adapted process in $\mathcal{L}^2(\mu_T\times\mathbb{P})$. Then the Itô integral $I_t(X_\cdot)$, $t\in[0,T]$ can be chosen to have continuous sample paths.*[4]

Proof. As usual, we choose a sequence of simple approximations X^n_t. Then $I_t(X^n_\cdot)$ is a martingale with continuous sample paths, and so is $M^n_t = I_t(X^n_\cdot) - I_t(X^{n-1}_\cdot)$. By Jensen's inequality $\mathbb{E}((M^n_t)^2|\mathcal{F}_s) \geq (\mathbb{E}(M^n_t|\mathcal{F}_s))^2 = (M^n_s)^2$ for $s\leq t$, so $(M^n_t)^2$ is a submartingale. Define the stopping time $\tau = \inf\{t\in[0,T] : |M^n_t|\geq\varepsilon\}$. Then

$$\mathbb{P}\left(\sup_{t\in[0,T]}|M^n_t|>\varepsilon\right) \leq \mathbb{P}\left((M^n_\tau)^2\geq\varepsilon^2\right) \leq \varepsilon^{-2}\,\mathbb{E}\left((M^n_\tau)^2\right) \leq \varepsilon^{-2}\,\mathbb{E}\left((M^n_T)^2\right),$$

where we have used continuity of the sample paths, Chebyshev's inequality, and the submartingale property. In particular, we obtain the estimate

$$\mathbb{P}\left(\sup_{t\in[0,T]}\left|\int_0^t (X^n_s - X^{n-1}_s)\,dW_s\right| > \frac{1}{n^2}\right) \leq n^4\,\mathbb{E}\left[\int_0^T (X^n_s - X^{n-1}_s)^2\,ds\right].$$

But we may assume that $\|X^n_\cdot - X^{n-1}_\cdot\|_{2,\mu_T\times\mathbb{P}} \leq 2^{-n}$; if this is not the case, we can always choose a subsequence $m(n)\nearrow\infty$ such that $\|X^{m(n)}_\cdot - X^{m(n-1)}_\cdot\|_{2,\mu_T\times\mathbb{P}}\leq 2^{-n}$. Thus we find, proceding with a suitable subsequence if necessary, that

$$\sum_{n=2}^\infty \mathbb{P}\left(\sup_{t\in[0,T]}\left|\int_0^t (X^n_s - X^{n-1}_s)\,dW_s\right| > \frac{1}{n^2}\right) < \infty.$$

But then it follows, using the Borel-Cantelli lemma, that

$$\sum_{n=2}^\infty \sup_{t\in[0,T]}\left|\int_0^t (X^n_s - X^{n-1}_s)\,dW_s\right| < \infty \quad \text{a.s.},$$

and hence $I_t(X^n_\cdot)$ a.s. converges uniformly to some process H_t with continuous sample paths. As the discontinuous paths live in a null set, we may set them to zero without inflicting any harm. It remains to show that for every $t\in[0,T]$, the random variable H_t is the limit in $\mathcal{L}^2(\mathbb{P})$ of $I_t(X^n_\cdot)$, i.e., that H_t coincides with the definition of the elementary Itô integral for every time t. But as $I_t(X^n_\cdot)\to I_t(X_\cdot)$ in $\mathcal{L}^2(\mathbb{P})$ and $I_t(X^n_\cdot)\to H_t$ a.s., we find that

$$\mathbb{E}((H_t - I_t(X_\cdot))^2) = \mathbb{E}\Big(\liminf_{n\to\infty}(I_t(X^n_\cdot) - I_t(X_\cdot))^2\Big) \leq \liminf_{n\to\infty}\mathbb{E}((I_t(X^n_\cdot) - I_t(X_\cdot))^2) = 0$$

where we have used Fatou's lemma. Hence $H_t = I_t(X_\cdot)$ a.s., and we are done. □

[4] An immediate consequence is that any version of the Itô integral has a continuous modification.

Localization

We would like to define the Itô integral as a continuous process on the entire interval $[0,\infty[$. Which integrands can we do this for? The most straightforward idea would be to require the integrands to be $\mathcal{F}_t$-adapted processes in $\mathcal{L}^2(\mu\times\mathbb{P})$, where μ is the Lebesgue measure on $[0,\infty[$. This would indeed be necessary if we wish to define

$$I_\infty(X_\cdot) = \int_0^\infty X_t\,dW_t,$$

but this is not our goal: we only wish to define the integral as a stochastic process $I_t(X_\cdot)$ for every finite time $t\in[0,\infty[$, and we do not necessarily care whether $I_\infty(X_\cdot)$ actually exists. Hence the condition $X_\cdot\in\mathcal{L}^2(\mu\times\mathbb{P})$ seems excessively restrictive.

To weaken this condition, we use a trick called *localization*. This is not a very deep idea. To define $I_t(X_\cdot)$ on $[0,\infty[$, it suffices that we can define it on every interval $[0,T]$: after all, to compute X_t for fixed t, we can simply choose $T\ge t$ and proceed with the construction in the previous sections. Hence it should suffice to require that $X_{[0,T]}\in\mathcal{L}^2(\mu_T\times\mathbb{P})$ for every $T<\infty$, i.e., that $X_\cdot\in\bigcap_{T<\infty}\mathcal{L}^2(\mu_T\times\mathbb{P})$, a much weaker condition than $X_\cdot\in\mathcal{L}^2(\mu\times\mathbb{P})$! This is called localization, because we have taken a *global* construction on $[0,T]$—in the previous section we defined the sample paths of $I_t(X_\cdot)$ on all of $[0,T]$ simultaneously—and applied it *locally* to every subinterval $[0,T]\subset[0,\infty[$. The advantage is that the integrands need not be square integrable on $[0,\infty[\times\Omega$; they only need to be *locally* square integrable, i.e., square integrable when restricted to any bounded set of times $[0,T]$.

It is not immediately obvious that this procedure is consistent, however. We have to verify that our definition of $I_t(X_\cdot)$ does not depend on which $T>t$ we choose for its construction; if the definition does depend on T, then our localization procedure is ambiguous! Fortunately, the local property of the Itô integral is easy to verify.

Lemma 4.2.8. *For any $\mathcal{F}_t$-adapted process $X_\cdot\in\bigcap_{T<\infty}\mathcal{L}^2(\mu_T\times\mathbb{P})$, we can define uniquely the Itô integral $I_t(X_\cdot)$ as an $\mathcal{F}_t$-adapted stochastic process on $[0,\infty[$ with continuous sample paths.*

Proof. For any finite time T, we can construct the Itô integral of $X_{[0,T]}$ as a stochastic process on $[0,T]$ with continuous sample paths (by lemma 4.2.7), and it is clear from the construction that $X_t^T(X_\cdot)$ is $\mathcal{F}_t$-adapted. Let us call the process thus constructed $I_t^T(X_\cdot)$. We would like to prove that for fixed T, $\mathbb{P}(I_s^t(X_\cdot)=I_s^T(X_\cdot)\text{ for all }s\le t\le T)=1$. But this is immediate from the definition when $X_\cdot$ is a simple integrand, and follows for the general case by choosing the same approximating sequence X_t^n, defined on $[0,T]$, to define both $I_s^t(X_\cdot)$ and $I_s^T(X_\cdot)$. Finally, as this holds for any $T\in\mathbb{N}$, we find $\mathbb{P}(I_s^t(X_\cdot)=I_s^T(X_\cdot)\text{ for all }s\le t\le T<\infty)=1$, so that $I_t(X_\cdot)$ is unambiguously defined by setting $I_t(X_\cdot)=I_t^T(X_\cdot)$ for any $T\ge t$. □

Somewhat surprisingly, the simple concept of localization can be used to extend the class of integrable processes even beyond the locally square integrable processes. To see this, we first need to investigate how the Itô integral behaves under stopping.

Lemma 4.2.9. *Let X_t be an $\mathcal{F}_t$-adapted process in $\bigcap_{T<\infty}\mathcal{L}^2(\mu_T\times\mathbb{P})$, and let τ be an $\mathcal{F}_t$-stopping time. Then $I_{t\wedge\tau}(X_\cdot)=I_t(X_\cdot I_{\cdot<\tau})$.*

Proof. As τ is a stopping time, $I_{t<\tau}$ is $\mathcal{F}_t$-adapted, and hence $X_t I_{t<\tau}$ is $\mathcal{F}_t$-adapted and in $\bigcap_{T<\infty} \mathcal{L}^2(\mu_T \times \mathbb{P})$. Hence the integral in the statement of the lemma exists. To prove the result, fix some interval $[0,T]$ and choose a sequence X_t^n of simple processes on $[0,T]$ that converge to X_t fast enough. Let us suppose additionally that $\tau \leq T$ a.s. Define the random time τ^n to be the value of τ rounded *upwards* to the earliest jump time of X_t^n that is larger or equal to τ. The times τ^n are still stopping times, and thus $X_t^n I_{t<\tau^n}$ is a sequence of $\mathcal{F}_t$-adapted simple approximations that converges to $X_t I_{t\leq\tau}$. But for the simple integrands, you can verify immediately that $I_T(X_\cdot^n I_{\cdot<\tau^n}) = I_{\tau^n}(X_\cdot^n)$, and hence we find that $I_T(X_\cdot I_{\cdot<\tau}) = I_\tau(X_\cdot)$ by letting $n \to \infty$ and using continuity of the sample paths. When τ is not bounded by T, we simply apply the above procedure to the bounded stopping time $\tau \wedge T$. Finally, we have only proved the statement for every T separately, so you might worry that the processes $I_{T\wedge\tau}(X_\cdot)$ and $I_T(X_\cdot I_{\cdot<\tau})$ are only modificiations of each other. But both these processes have continuous sample paths, and modifications with continuous paths are indistinguishable. □

We would like to weaken the requirement $X_\cdot \in \bigcap_{T<\infty} \mathcal{L}^2(\mu_T \times \mathbb{P})$, i.e.,

$$\mathbb{E}\left[\int_0^T X_t^2\, dt\right] < \infty \quad \text{for all } T < \infty.$$

Recall that in this case, localization consisted of defining $I_t(X_\cdot)$ as $I_t^T(X_\cdot)$ (the integral constructed in $\mathcal{L}^2(\mu_T \times \mathbb{P})$) for T large enough, and lemma 4.2.8 guarantees that the definition does not depend on the particular choice of T. Now suppose that instead of the above condition, we are in a situation where

$$\mathbb{E}\left[\int_0^{\tau_n} X_t^2\, dt\right] < \infty \quad \text{for all } n \in \mathbb{N},$$

where $\tau_n \nearrow \infty$ is some sequence of $\mathcal{F}_t$-stopping times. Then $\{\tau_n\}$ is called a *localizing sequence* for X_t. Even though X_t need not be in $\mathcal{L}^2(\mu_T \times \mathbb{P})$ for any T, it is clear from this definition that the process $X_t I_{t<\tau_n}$ is in $\bigcap_{T<\infty} \mathcal{L}^2(\mu_T \times \mathbb{P})$, and hence $I_t(X_\cdot I_{\cdot<\tau_n})$ is well defined by our earlier efforts. In view of lemma 4.2.9, we should thus consider defining the Itô integral $I_t(X_\cdot) = I_t(X_\cdot I_{\cdot<\tau_n})$ for all $t \leq \tau_n$, and as $\tau_n \nearrow \infty$ we can always choose n large enough so that this definition makes sense. This is precisely the same idea as our previous exercise in localization, except that we are now allowing our localization intervals $[0,\tau_n]$ to be random.

Once again, the big question is whether the definition of $I_t(X_\cdot)$ depends on the choice of n, or indeed on the choice of localizing sequence $\{\tau_n\}$ (as for any integrand X_t, there could be many localizing sequences that work!)

Lemma 4.2.10. *Let X_t be an $\mathcal{F}_t$-adapted process which admits a localizing sequence τ_n. Then $I_t(X_\cdot)$ is uniquely defined as an $\mathcal{F}_t$-adapted stochastic process on $[0,\infty[$ with continuous sample paths and is independent of the choice of localizing sequence.*

Proof. First, we claim that for every $m > n$, we have $I_t(X_\cdot I_{\cdot<\tau_n}) = I_t(X_\cdot I_{\cdot<\tau_m})$ for all $t < \tau_n$. But as $\tau_m \geq \tau_n$ by assumption, clearly $I_{t<\tau_n} I_{t<\tau_m} = I_{t<\tau_n}$, so by lemma 4.2.9 we find that $I_t(X_\cdot I_{\cdot<\tau_n}) = I_{t\wedge\tau_n}(X_\cdot I_{\cdot<\tau_m})$ which establishes the claim. Hence we can unambiguously define $I_t(X_\cdot)$ as $I_t(X_\cdot I_{\cdot<\tau_n})$ for all $t < \tau_n$. Moreover, the integral thus defined is clearly $\mathcal{F}_t$-adapted and has continuous sample paths.

It remains to show that the definition thus obtained does not depend on the choice of localizing sequence. To see this, let τ'_n be another localizing sequence for X_t, and denote by $I'_t(X_\cdot)$ the Itô integral constructed using this sequence. Introduce also the stopping times $\sigma_n = \tau_n \wedge \tau'_n$, and note that this forms another localizing sequence. Denote by $J_t(X_\cdot)$ the Itô integral constructed using σ_n. But then, by the same argument as above, $J_t(X_\cdot) = I_t(X_\cdot)$ and $J_t(X_\cdot) = I'_t(X_\cdot)$ for all $t \le \sigma_n$, and hence $I'_t(X_\cdot) = I_t(X_\cdot)$ for all t. □

How does this help us? There is a natural class of integrands—much larger than $\bigcap_{T<\infty} \mathcal{L}^2(\mu_T \times \mathbb{P})$—whose elements admit a localizing sequence. We only need to require that the integrand X_t is $\mathcal{F}_t$-adapted and satisfies

$$A_T(X_\cdot) = \int_0^T X_t^2 \, dt < \infty \quad \text{a.s.} \quad \text{for all } T < \infty.$$

To exhibit a localizing sequence, consider the stopping times $\tau_n = \inf\{t \le n : A_t(X_\cdot) \ge n\}$. Then the condition on our integrand guarantees that $\tau_n \nearrow \infty$ a.s., and

$$\int_0^{\tau_n} X_t^2 \, dt \le n \quad \text{a.s.} \quad \text{for all } n \in \mathbb{N}.$$

Taking the expectation, we see that evidently τ_n is a localizing sequence of X_t. Let us finally summarize what we have learned.

Definition 4.2.11 (Itô integral). Let X_t be any $\mathcal{F}_t$-adapted stochastic process with

$$\mathbb{P}\left[\int_0^T X_t^2 \, dt < \infty\right] = 1 \quad \text{for all } T < \infty.$$

Then the Itô integral

$$I_t(X_\cdot) = \int_0^t X_s \, dW_s$$

is uniquely defined, by localization and the choice of a continuous modification, as an $\mathcal{F}_t$-adapted stochastic process on $[0, \infty[$ with continuous sample paths.

Remark 4.2.12. There is another approach to this definition. We can prove an *in probability* version of the Itô isometry to replace the $\mathcal{L}^2$ version used in the construction of the elementary integral, see [Fri75, lemma 4.2.3] or [LS01a, lemma 4.6]. Using this weaker result, we can then prove that for any integrand which satisfies the condition in the definition, we can define the Itô integral as the limit in probability of a sequence of simple integrals. The integral thus defined coincides with our definition through localization, but provides a complementary view on its construction.

Remark 4.2.13. The generalization to the class of integrands in definition 4.2.11 will make the stochastic calculus much more transparent. If you care about generality for its own sake, however, you might be interested to know that the current conditions can not be reasonably relaxed; see [McK69, section 2.5, problem 1].

4.3 Some elementary properties

We have done some heavy work in constructing a solid Itô integral; we will pick the fruits of this labor throughout the rest of this course. Before we move on, let us spend a few relaxing moments proving some of the simplest properties of the Itô integral.

Lemma 4.3.1 (Linearity). *Let X_t and Y_t be Itô integrable processes, and let $\alpha, \beta \in \mathbb{R}$. Then $I_t(\alpha X_\cdot + \beta Y_\cdot) = \alpha\, I_t(X_\cdot) + \beta\, I_t(Y_\cdot)$.*

Proof. This clearly holds for simple integrands in $\mathcal{L}^2(\mu_T \times \mathbb{P})$, and hence follows for any locally square integrable integrand by taking limits. It remains to note that we can choose a common localizing sequence for X_t and Y_t: choosing a localizing sequence σ_n and τ_n for X_t and Y_t, respectively, the sequence $\sigma_n \wedge \tau_n$ is localizing for both. Hence we can directly extend to any admissible integrands X_t and Y_t by localization with respect to $\sigma_n \wedge \tau_n$. □

Lemma 4.3.2. *Let X_t be Itô integrable and let τ be an $\mathcal{F}_t$-stopping time. Then*

$$\int_0^{t\wedge\tau} X_s\, dW_s = \int_0^t X_s\, I_{s<\tau}\, dW_s.$$

Proof. For integrands in $\bigcap_{T<\infty} \mathcal{L}^2(\mu_T \times \mathbb{P})$, this follows from lemma 4.2.9. For the general case, let σ_n be a localizing sequence. Then by definition, $I_{t\wedge\tau}(X_\cdot) = I_{t\wedge\tau}(X_\cdot I_{\cdot<\sigma_n})$ for $t < \sigma_n$, and using lemma 4.2.9 we find $I_{t\wedge\tau}(X_\cdot I_{\cdot<\sigma_n}) = I_t(X_\cdot I_{\cdot<\tau} I_{\cdot<\sigma_n})$. But σ_n is clearly also a localizing sequence for $X_t I_{t<\tau}$, so the result follows by localization. □

When the integrand is in $\bigcap_{T<\infty} \mathcal{L}^2(\mu_T \times \mathbb{P})$, the Itô integral inherits the elementary properties of the simple integrals. This is *very* convenient in computations.

Lemma 4.3.3. *Let $X_\cdot \in \bigcap_{T<\infty} \mathcal{L}^2(\mu_T \times \mathbb{P})$. Then for any $T < \infty$*

$$\mathbb{E}\left[\int_0^T X_t\, dW_t\right] = 0, \qquad \mathbb{E}\left[\left(\int_0^T X_t\, dW_t\right)^2\right] = \mathbb{E}\left[\int_0^T X_t^2\, dt\right],$$

and moreover $I_t(X_\cdot)$ is an $\mathcal{F}_t$-martingale.

Proof. All the statements of this lemma hold when X_t is a simple integrand in $\mathcal{L}^2(\mu_T \times \mathbb{P})$, as we have seen in the construction of the integral. The result follows directly, as $Y_n \to Y$ in $\mathcal{L}^2$ implies that $\mathbb{E}(Y_n) \to \mathbb{E}(Y)$, $\mathbb{E}(Y_n|\mathcal{F}) \to \mathbb{E}(Y|\mathcal{F})$ in $\mathcal{L}^2$, and $\mathbb{E}(Y_n^2) \to \mathbb{E}(Y^2)$ (why?). □

As an immediate corollary, we find that if a sequence of integrands converges in $\mathcal{L}^2(\mu_T \times \mathbb{P})$, then so do their Itô integrals. We already used this property for simple integrands, but the fact that this holds generally is often useful. This should be more or less trivial by now, so there is no need to provide a proof.

Corollary 4.3.4. *If $X_\cdot^n \to X_\cdot$ in $\mathcal{L}^2(\mu_T \times \mathbb{P})$, then $I_t(X_\cdot^n) \to I_t(X_\cdot)$ in $\mathcal{L}^2(\mathbb{P})$. Moreover, if the convergence is fast enough, then $I_t(X_\cdot^n) \to I_t(X_\cdot)$ a.s.*

In the general case, i.e., when $X_\cdot \notin \bigcap_{T<\infty} \mathcal{L}^2(\mu_T \times \mathbb{P})$, the nice properties of lemma 4.3.3 are unfortunately not guaranteed. In fact, in the general case X_t may not even be in $\mathcal{L}^1(\mathbb{P})$, in which case its expectation need not be defined and the martingale property need not even make sense. However, there is a weakening of the martingale property that is especially suitable for stochastic integration.

Definition 4.3.5. An $\mathcal{F}_t$-measurable process X_t is called an *$\mathcal{F}_t$-local martingale* if there exists a sequence of $\mathcal{F}_t$-stopping times $\tau_n \nearrow \infty$ such that $X_{t\wedge\tau_n}$ is a martingale for every n. The sequence τ_n is called a *reducing sequence* for X_t.

Lemma 4.3.6. *Any Itô integral $I_t(X_\cdot)$ is a local martingale.*

Proof. Any localizing sequence for X_t is a reducing sequence for $I_t(X_\cdot)$. □

Remark 4.3.7. The local martingale property is fundamental in the general theory of stochastic integration, and is intimately related with the notion of localization. However, lemma 4.3.3 shows that the integrands in $\bigcap_{T<\infty} \mathcal{L}^2(\mu_T \times \mathbb{P})$ behave much nicer in computations, at least where expectations are involved, than their more general localized counterparts. When applying the Itô calculus, to be developed next, localization will allow us to manipulate the stochastic integrals very easily; but at the end of the day we will still need to prove separately that the resulting integrands are in $\bigcap_{T<\infty} \mathcal{L}^2(\mu_T \times \mathbb{P})$ if we wish to calculate the expectation.

4.4 The Itô calculus

Perhaps the most important topic in stochastic integration is the associated *calculus*, which gives us transparent tools to manipulate Itô integrals and stochastic differential equations. You know the deterministic counterpart of this idea very well. For example, if f is C^2 (twice continuously differentiable), then

$$X(t) = X(0) + \int_0^t Y(t)\,dt \quad \Longrightarrow \quad f(X(t)) = f(X(0)) + \int_0^t \frac{df}{dx}(X_t)\,Y(t)\,dt.$$

This in itself is not a trivial result; it requires some amount of analytic machinery to prove! However, after the analysis has been done once, we no longer need to do any work to apply the result; all we have to remember is a simple *rule* on how integrals and derivatives transform, and every time we apply this rule a whole bunch of analysis goes on under the hood. The rule is so simple, in fact, that it can be taught to high school students with no background in analysis whatsoever! Our goal here is to find a similar rule for the Itô integral—a different one than the usual rule, but just as easy to apply—which pushes almost all of the difficult analysis out of view.

We will work in a rather general setup, in view of applications to (multidimensional) stochastic differential equations. Let us work on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\in[0,\infty[}, \mathbb{P})$ on which we have defined an m-dimensional $\mathcal{F}_t$-Wiener process $W_t = (W_t^1, \ldots, W_t^m)$ (i.e., W_t^i are independent $\mathcal{F}_t$-Wiener processes). We consider $\mathcal{F}_t$-adapted processes $X^1, \ldots, X^n$ of the form

$$X_t^i = X_0^i + \int_0^t F_s^i\,ds + \sum_{j=1}^m \int_0^t G_s^{ij}\,dW_s^j,$$

where F_s^i, G_s^{ij} are $\mathcal{F}_t$-progressively measurable processes that satisfy

$$\int_0^t |F_s^i|\,ds < \infty, \quad \int_0^t (G_s^{ij})^2\,ds < \infty \quad \text{a.s.} \qquad \forall\, t < \infty, \quad \forall\, i, j.$$

We call $X_t = (X_t^1, \ldots, X_t^n)$ an *n-dimensional Itô process*.

Definition 4.4.1 (Itô process). A process $X_t = (X_t^1, \ldots, X_t^n)$ satisfying the above conditions is called an *n-dimensional Itô process*. It is also denoted as

$$X_t = X_0 + \int_0^t F_s\,ds + \int_0^t G_s\,dW_s.$$

The goal of this section is to prove the following theorem.

Theorem 4.4.2 (Itô rule). *Let $u : [0, \infty[\times\mathbb{R}^n \to \mathbb{R}$, be a function such that $u(t, x)$ is C^1 with respect to t and C^2 with respect to x. Then $u(t, X_t)$ is an Itô process itself:*

$$u(t, X_t) = u(0, X_0) + \sum_{i=1}^n \sum_{k=1}^m \int_0^t u_i(s, X_s)\, G_s^{ik}\, dW_s^k$$
$$+ \int_0^t \left\{ u'(s, X_s) + \sum_{i=1}^n u_i(s, X_s)\, F_s^i + \frac{1}{2} \sum_{i,j=1}^n \sum_{k=1}^m u_{ij}(s, X_s)\, G_s^{ik}\, G_s^{jk} \right\} ds,$$

where we have written $u'(t, x) = \partial u(t, x)/\partial t$ and $u_i(t, x) = \partial u(t, x)/\partial x^i$.

Remark 4.4.3 (Itô differentials). We will often use another notation for the Itô process, particularly when dealing with stochastic differential equations:

$$dX_t = F_t\,dt + G_t\,dW_t.$$

Though this expression is reminiscent of derivatives in ordinary calculus, this is just suggestive notation for the expression for X_t in integrated form, as in definition 4.4.1. However, it takes up much less room, and allows for very quick symbolic computations. To see how, let us rewrite the Itô rule in symbolic form:

$$du(t, X_t) = u'(t, X_t)\,dt + \partial u(t, X_t)\,dX_t + \frac{1}{2}\,\mathrm{Tr}[\partial^2 u(t, X_t)\,dX_t (dX_t)^*],$$

where $\partial u(t, x)$ denotes the row vector with elements $u_i(t, x)$, $\partial^2 u(t, x)$ is the matrix with entries $u_{ij}(t, x)$, and $dX_t^i\,dX_t^j$ is calculated according to the following *Itô table:*

	dt	dW_t^j
dt	0	0
dW_t^i	0	$\delta_{ij}\,dt$

For example, if (in one dimension) $dX_t = F_t\,dt + G_t^1\,dW_t^1 + G_t^2\,dW_t^2$, then $(dX_t)^2 = \{(G_t^1)^2 + (G_t^2)^2\}\,dt$. You should inspect the symbolic expression of the Itô rule carefully and convince yourself that it does indeed coincide with theorem 4.4.2.

You can now easily see how extraordinarily simple Itô's rule really is. If we write our processes in terms of the "differentials" dX_t, dW_t^i, etc., then Itô's rule reduces to applying some easy to remember multiplication rules to the differentials. This is highly reminiscent of the chain rule in ordinary calculus—the first two terms of the Itô rule *are* the ordinary chain rule, and Itô's rule does indeed reduce to the chain rule when $G_t^{ij} = 0$ (as it should!) When stochastic integrals are present, we evidently need to take a second-order term into account as well. Once you get used to pushing around the various symbols in the right way, applying Itô's rule will be no more difficult than calculating derivatives in your high school calculus class.

Important example 4.4.4. Let X_t^1, X_t^2 be two one-dimensional Itô processes, and consider the function $u(t, x^1, x^2) = x^1 x^2$. Then $u \in C^2$, and so by Itô's rule

$$X_t^1 X_t^2 = X_0^1 X_0^2 + \sum_{k=1}^{m} \int_0^t \{X_s^1 G_s^{2k} + X_s^2 G_s^{1k}\}\, dW_s^k + \int_0^t \left\{ X_s^1 F_s^2 + X_s^2 F_s^1 + \sum_{k=1}^{m} G_s^{1k} G_s^{2k} \right\} ds.$$

In differential form, we find the *product rule* $d(X_t^1 X_t^2) = X_t^1\, dX_t^2 + X_t^2\, dX_t^1 + dX_t^1\, dX_t^2$. In particular, we see that the class of Itô processes is closed under multiplication. Moreover, the class of Itô processes is trivially closed under the formation of linear combinations, so apparently the Itô processes form an *algebra*.[5]

Let us now proceed to the proof of Itô's rule. The proof goes a little fast at times; if we would work out every minor point in the goriest detail, the proof would be many pages longer. You can fill in the details yourself without too much effort, or look them up in one of the references mentioned in section 4.7.

Proof of theorem 4.4.2. Until further notice, we consider the case where the integrands F_t^i and G_t^{ij} are bounded, simple, and $\mathcal{F}_t$-adapted. We also assume that $u(t, x)$ is independent of t and that all first and second derivatives of u are uniformly bounded. Once we have proved the Itô rule for this special case, we will generalize to obtain the full Itô rule.

The key, of course, is Taylor's formula [Apo69, theorem 9.4]:

$$u(x) = u(x_0) + \partial u(x_0)\,(x - x_0) + \frac{1}{2}\,(x - x_0)^*\,\partial^2 u(x_0)\,(x - x_0) + \|x - x_0\|^2\, E(x, x_0),$$

where $E(x, x_0)$ is uniformly bounded and $E(x, x_0) \to 0$ as $\|x - x_0\| \to 0$. As F_t^i and G_t^{ij} are simple integrands, we may assume that they all have the same jump times t_k, $k = 1, \ldots, N$ (otherwise we can join all the jump times into one sequence, and proceed with that). Write

$$u(X_t) = u(X_0) + \sum_{k=0}^{N} (u(X_{t_{k+1}}) - u(X_{t_k})),$$

where we use the convention $t_0 = 0$ and $t_{N+1} = t$. We are going to deal with every term of this sum separately. Let us thus fix some k, and define $s_\ell^p = t_k + \ell 2^{-p}(t_{k+1} - t_k)$. Then

$$u(X_{t_{k+1}}) - u(X_{t_k}) = \sum_{\ell=1}^{2^p} (u(X_{s_\ell^p}) - u(X_{s_{\ell-1}^p}))$$

[5] This would not be true without localization.

for any $p \in \mathbb{N}$. Note that as F_t and G_t are constant in the interval $t_{k+1} - t_k$, we have

$$X_{s_\ell^p} - X_{s_{\ell-1}^p} = F_{t_k}\, 2^{-p}\Delta + G_{t_k}\,(W_{s_\ell^p} - W_{s_{\ell-1}^p}).$$

Thus applying Taylor's formula to $u(X_{s_\ell^p}) - u(X_{s_{\ell-1}^p})$, we get

$$\begin{aligned}
&u(X_{t_{k+1}}) - u(X_{t_k}) = \\
&\quad \sum_{\ell=1}^{2^p} \partial u(X_{s_{\ell-1}^p})\,(F_{t_k}\, 2^{-p}\Delta + G_{t_k}\,(W_{s_\ell^p} - W_{s_{\ell-1}^p}) && (4.4.1)\\
&\quad + \frac{1}{2}\sum_{\ell=1}^{2^p} (G_{t_k}\,(W_{s_\ell^p} - W_{s_{\ell-1}^p}))^*\,\partial^2 u(X_{s_{\ell-1}^p})\,(G_{t_k}\,(W_{s_\ell^p} - W_{s_{\ell-1}^p})) && (4.4.2)\\
&\quad + \frac{1}{2}(F_{t_k}\, 2^{-p}\Delta)^* \sum_{\ell=1}^{2^p} \partial^2 u(X_{s_{\ell-1}^p})\,(F_{t_k}\, 2^{-p}\Delta) && (4.4.3)\\
&\quad + (F_{t_k}\, 2^{-p}\Delta)^* \sum_{\ell=1}^{2^p} \partial^2 u(X_{s_{\ell-1}^p})\,(G_{t_k}\,(W_{s_\ell^p} - W_{s_{\ell-1}^p})) && (4.4.4)\\
&\quad + \sum_{\ell=1}^{2^p} \|F_{t_k}\, 2^{-p}\Delta + G_{t_k}\,(W_{s_\ell^p} - W_{s_{\ell-1}^p})\|^2\, E(X_{s_\ell^p}, X_{s_{\ell-1}^p}). && (4.4.5)
\end{aligned}$$

We now let $p \to \infty$ and look whether the various terms on the right converge in $\mathcal{L}^2(\mathbb{P})$.

Consider first the terms (4.4.3) and (4.4.4). As $\partial^2 u(X_s)$ is bounded and has continuous sample paths, the sums in (4.4.3) and (4.4.4) converge in $\mathcal{L}^2(\mathbb{P})$ to a time integral and an Itô integral, respectively. But both terms are premultiplied by 2^{-p}, so we conclude that these terms converge to zero. Next, note that the term (4.4.5) can be estimated as

$$\sup_\ell |E(X_{s_\ell^p}, X_{s_{\ell-1}^p})| \left\{ \Delta^2 \|F_{t_k}\|^2 + \|G_{t_k}\|^2 \sum_{\ell=1}^{2^p} \|W_{s_\ell^p} - W_{s_{\ell-1}^p}\|^2 \right\}.$$

But the sum converges in $\mathcal{L}^2(\mathbb{P})$ to the quadratic variation $m\Delta$, while the supremum term is uniformly bounded and converges a.s. to zero. Hence the entire term converges to zero in $\mathcal{L}^2(\mathbb{P})$. Next, note that the term (4.4.1) converges in $\mathcal{L}^2(\mathbb{P})$ to

$$\int_{t_k}^{t_{k+1}} \partial u(X_r)\, F_{t_k}\, dr + \int_{t_k}^{t_{k+1}} \partial u(X_r)\, G_{t_k}\, dW_r.$$

It remains to investigate the term (4.4.2). We claim that this term converges in $\mathcal{L}^2(\mathbb{P})$ to

$$\frac{1}{2}\int_{t_k}^{t_{k+1}} \mathrm{Tr}[\partial^2 u(X_r)\, G_{t_k} (G_{t_k})^*]\, dr.$$

But this calculation is almost identical to the calculation of the quadratic variation of the Wiener process in the proof of lemma 3.1.11, so we will leave it as an exercise.

Finally, summing over all k, we find that Itô's rule is indeed satisfied in the case that F_t^i and G_t^{ij} are $\mathcal{F}_t$-adapted bounded simple integrands, $u(t,x)$ is independent of t and has uniformly bounded first and second derivatives. We now need to generalize this statement.

First, suppose that $G_t^{ik} \in \mathcal{L}^2(\mu_T \times \mathbb{P})$ for all i,k, and that F_t satisfies the general condition of the theorem. Then we can find a sequence of simple approximations to F_t, G_t which

converges fast enough, so that the simple Itô processes converge a.s. to X_t. Moreover, the Itô rule holds for each of these simple approximations. But as we have assumed that the derivatives of u are bounded and continuous, it is easy to see that the integrands obtained by applying the Itô rule to the simple approximations converge sufficiently fast to the integrands in the Itô rule applied to X_t. Taking the a.s. limit, we can conclude using corollary 4.3.4 that the Itô rule still holds when $G_t^{ik} \in \mathcal{L}^2(\mu_T \times \mathbb{P})$ and F_t satisfies the general condition of the theorem.

Our next job is to add time to the picture (but u is still bounded). If $u(t,x)$ is required to be C^2 in all variables, then this is trivial: $X_t^i = t$ is an Itô process, so we can always extend the dimension of our Itô process by one to include time in the picture. To allow u to be only C^1 in time, we can always find (e.g., by convolution) a sequence u^n of C^2 approximations to u so that u^n, u_i^n, u_{ij}^n and $(u^n)'$ converge uniformly on compact sets fast enough. The result follows by taking the limit. (Actually, if it happens to be the case for some i that $G^{im} = 0$ for all m, then we could similarly only require u to be C^1 in the variable x^i.)

It remains to weaken the requirement that G_t is square-integrable and that u has bounded derivatives. But we can solve both these problems simultaneously by localization. Indeed, choose a localizing sequence τ_n for X_t, and choose another sequence of stopping times $\sigma_n \nearrow \infty$ such that $u_i(t, X_t)$, $u_{ij}(t, X_t)$ and $u'(t, X_t)$ are bounded by n for all $t < \sigma_n$. Then $\tau_n \wedge \sigma_n$ is another localizing sequence, and we can apply Itô's rule to $X_{t \wedge \tau_n \wedge \sigma_n}$. We are done. □

Remark 4.4.5. Suppose that for all times t, the Itô process X_t a.s. takes values in some open set U. Then, using another localization trick, it suffices for Itô's theorem that $u(t,x)$ is C^1 in t and C^2 for $x \in U$. Proving this is a good exercise in localization.

For example, we will sometimes encounter an Itô process such that $X_t > 0$ a.s. for all t. We can then apply Itô's rule with $u(t,x) = \sqrt{x}$ or even $u(t,x) = x^{-1}$, even though these functions are not C^2 on all of $\mathbb{R}$ (but they are C^2 on $]0, \infty[$).

4.5 Girsanov's theorem

In this and the next section, we will discuss two fundamental theorems of stochastic analysis. In this section we discuss Girsanov's theorem, which tells us what happens to the Wiener process under a change of measure. This theorem has a huge number of applications, some of which we will encounter later on in this course. In order to lead up to the result, however, let us first consider a pair of illustrative discrete examples.

Example 4.5.1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space on which is defined a Gaussian random variable ξ with zero mean and unit variance, and let $a \in \mathbb{R}$ be an arbitrary (non-random) constant. We would like to find a new measure $\mathbb{Q}$ under which $\xi' = a + \xi$ is a Gaussian random variable with zero mean and unit variance. In other words, we would like the law of ξ' under $\mathbb{Q}$ to equal the law of ξ under $\mathbb{P}$.

Of course, this is only possible if the laws μ_ξ and $\mu_{\xi'}$ (under $\mathbb{P}$) of ξ and ξ', respectively, are absolutely continuous. But indeed this is the case: note that

$$\mu_\xi(A) = \int_A \frac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx, \qquad \mu_{\xi'}(A) = \int_A \frac{e^{-(x-a)^2/2}}{\sqrt{2\pi}}\,dx,$$

so evidently

$$\frac{d\mu_\xi}{d\mu_{\xi'}}(x) = e^{(x-a)^2/2 - x^2/2} = e^{-ax + a^2/2}.$$

It follows immediately that if we define $\mathbb{Q}$ by

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = e^{-a\xi' + a^2/2} = e^{-a\xi - a^2/2},$$

then the measure $\mathbb{Q}$ has the desired property.

Example 4.5.2. Now let $(\Omega, \mathcal{F}, \mathbb{P})$ carry a collection $\xi_1, \ldots, \xi_n$ of i.i.d. Gaussian random variables with zero mean and unit variance, and define the filtration $\mathcal{F}_k = \sigma\{\xi_1, \ldots, \xi_k\}$. Let $a_1, \ldots, a_n$ be a *predictable* process, i.e., a_k is an $\mathcal{F}_{k-1}$-measurable random variable. We would like to find a measure $\mathbb{Q}$ under which $\xi'_k = a_k + \xi_k$ are i.i.d. Gaussian random variables with zero mean and unit variance. In other words, we would like the law of the process ξ' under $\mathbb{Q}$ to equal the law of the process ξ under $\mathbb{P}$.

Let $f(x_1, \ldots, x_n)$ be any bounded measurable function. Under $\mathbb{P}$,

$$\mathbb{E}_{\mathbb{P}}(f(\xi'_1, \ldots, \xi'_n)) = \int_{\mathbb{R}^n} f(x_1 + a_1, \ldots, x_n + a_n(x_1, \ldots, x_{n-1})) \, \frac{e^{-(x_1^2 + \cdots + x_n^2)/2}}{(2\pi)^{n/2}} \, dx_1 \cdots dx_n,$$

where we have explicitly introduced the predictablity assumption by setting $a_k = a_k(\xi_1, \ldots, \xi_{k-1})$. Under $\mathbb{Q}$, on the other hand, we would like to have

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}(f(\xi'_1, \ldots, \xi'_n)) &= \int_{\mathbb{R}^n} f(x'_1, \ldots, x'_n) \, \frac{e^{-((x'_1)^2 + \cdots + (x'_n)^2)/2}}{(2\pi)^{n/2}} \, dx'_1 \cdots dx'_n \\ &= \int_{\mathbb{R}^n} f(x_1 + a_1, \ldots, x_n + a_n(x_1, \ldots, x_{n-1})) \\ &\qquad \times \frac{e^{-((x_1+a_1)^2 + \cdots + (x_n + a_n(x_1, \ldots, x_{n-1}))^2)/2}}{(2\pi)^{n/2}} \, dx_1 \cdots dx_n,\end{aligned}$$

where we have made a change of variables.[6] Thus evidently we should set

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = \frac{e^{-((\xi_1+a_1)^2 + \cdots + (\xi_n + a_n)^2)/2}}{e^{-(\xi_1^2 + \cdots + \xi_n^2)/2}} = \exp\left[\sum_{k=1}^n \left(-a_k \xi_k - \frac{1}{2} a_k^2\right)\right],$$

which is almost the same as in the previous example. (You should verify that this does not give the right answer if a_k is not assumed to be predictable!)

Apparently we can "add" to a sequence of i.i.d. Gaussian random variables an arbitrary predictable process simply by changing to an absolutely continuous probability measure. This can be very convenient. Many problems, random and non-random, can be simplified by making a suitable change of coordinates (using, e.g., Itô's rule). But here we have another tool at our disposal which is purely probabilistic in nature: we can try to simplify our problem by changing to a more convenient probability measure. We will see this idea in action, for example, when we discuss filtering.

[6] If you are unconvinced, assume first that f and $a_1, \ldots, a_n$ are smooth functions of $x_1, \ldots, x_n$, apply the usual change of variables formula, and then extend to arbitrary f and $a_1, \ldots, a_n$ by approximation.

With this bit of discrete time intuition under our belt, the Girsanov theorem should come as no great surprise. Indeed, it is simply the appropriate extension of the previous example, with a Wiener process replacing the i.i.d. sequence ξ_k.

Theorem 4.5.3 (Girsanov). *Let W_t be an m-dimensional $\mathcal{F}_t$-Wiener process on the probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\in[0,T]}, \mathbb{P})$, and let X_t be an Itô process of the form*

$$X_t = \int_0^t F_s\, ds + W_t, \qquad t \in [0,T].$$

Suppose furthermore that F_t is Itô integrable, and define

$$\Lambda = \exp\left[-\int_0^T (F_s)^* dW_s - \frac{1}{2}\int_0^T \|F_s\|^2\, ds\right]$$

($(F_s)^* dW_s = F_s^1\, dW_s^1 + \cdots + F_s^n\, dW_s^n$). If Novikov's condition

$$\mathbb{E}_{\mathbb{P}}\left[\exp\left(\frac{1}{2}\int_0^T \|F_s\|^2\, ds\right)\right] < \infty$$

is satisfied, then $\{X_t\}_{t\in[0,T]}$ is an $\mathcal{F}_t$-Wiener process under $\mathbb{Q}(A) = \mathbb{E}_{\mathbb{P}}(\Lambda I_A)$.

Remark 4.5.4. Instead of Novikov's condition, the weaker condition $\mathbb{E}_{\mathbb{P}}(\Lambda) = 1$ (which is implied by Novikov's condition, see theorem 4.5.8 below) is sufficient for the Girsanov theorem to hold; see, e.g., [Fri75, chapter 7]. Clearly this condition is necessary for Λ to be a Radon-Nikodym derivative—if F_t is too wild a process, we are in trouble! The Girsanov theorem says that the condition is also sufficient. However, this condition is often not so easy to check; Novikov's condition is by far the most useful sufficient condition in practice, and is relatively easy to verify in many cases.

Let us prove the Girsanov theorem. Some supporting lemmas of a technical nature are postponed until after the main portion of the proof.

Proof. Define the $\mathcal{F}_t$-adapted process

$$\Lambda_t = \exp\left[-\int_0^t (F_s)^* dW_s - \frac{1}{2}\int_0^t \|F_s\|^2\, ds\right], \qquad \Lambda_t = 1 - \int_0^t \Lambda_s\, (F_s)^* dW_s,$$

where we have applied Itô's rule to obtain the expression on the right. Hence Λ_t is a local martingale. In fact, Novikov's condition implies that Λ_t is a true martingale: see theorem 4.5.8.

Let us assume until further notice that F_t is bounded; we will generalize later. We wish to prove that X_t is an m-dimensional $\mathcal{F}_t$-Wiener process under $\mathbb{Q}$. Clearly X_t has continuous sample paths, so it suffices to verify its finite dimensional distributions. In fact, by a simple induction argument, it suffices to prove that for any $\mathcal{F}_s$-measurable bounded random variable Z, the increment $X_t - X_s$ ($t > s$) is an m-dimensional Gaussian random variable with zero mean and covariance matrix $(t-s)I$ (I is the identity matrix), independent of Z. We will do this as in the proof of theorem 3.2.5 by calculating the characteristic function. Given $\alpha \in \mathbb{R}^m$

and $\beta \in \mathbb{R}$, let us perform the following sequence of simple manipulations:[7]

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}(e^{i\alpha^*(X_t-X_s)+i\beta Z}) &= \mathbb{E}_{\mathbb{P}}(\Lambda_T e^{i\alpha^*(X_t-X_s)+i\beta Z}) = \mathbb{E}_{\mathbb{P}}(\mathbb{E}_{\mathbb{P}}(\Lambda_T|\mathcal{F}_t)\, e^{i\alpha^*(X_t-X_s)+i\beta Z}) \\ &= \mathbb{E}_{\mathbb{P}}(\Lambda_t e^{i\alpha^*(X_t-X_s)+i\beta Z}) = \mathbb{E}_{\mathbb{P}}(\Lambda_s e^{\int_s^t (i\alpha-F_r)^*\, dW_r + \int_s^t (i\alpha-F_r/2)^* F_r\, dr + i\beta Z}) \\ &= e^{-\|\alpha\|^2(t-s)/2}\mathbb{E}_{\mathbb{P}}(\Lambda_s e^{i\beta Z} e^{\int_s^t (i\alpha-F_r)^*\, dW_r - \frac{1}{2}\int_s^t \|i\alpha-F_r\|^2\, dr}).\end{aligned}$$

The essential claim is now that

$$\mathbb{E}_{\mathbb{P}}(e^{\int_s^t (i\alpha-F_r)^*\, dW_r - \frac{1}{2}\int_s^t \|i\alpha-F_r\|^2\, dr}|\mathcal{F}_s) = 1,$$

or, equivalently (why?), that the complex-valued stochastic process

$$M_t^\alpha = \exp\left[\int_0^t (i\alpha - F_r)^*\, dW_r - \frac{1}{2}\int_0^t \|i\alpha - F_r\|^2\, dr\right]$$

is a martingale. But by Itô's rule, we find that

$$M_t^\alpha = 1 + \int_0^t M_s^\alpha\, (i\alpha - F_s)^*\, dW_s,$$

and applying lemma 4.5.6 (use the boundedness of F_t) we find that the real and imaginary parts of $(i\alpha - F_s^i)M_s^\alpha$ are all in $\mathcal{L}^2(\mu_T \times \mathbb{P})$. Hence M_t^α is indeed a martingale. But then

$$\begin{aligned}&\mathbb{E}_{\mathbb{Q}}(e^{i\alpha^*(X_t-X_s)+i\beta Z}) \\ &\quad= e^{-\|\alpha\|^2(t-s)/2}\,\mathbb{E}_{\mathbb{P}}(\Lambda_s e^{i\beta Z}\,\mathbb{E}_{\mathbb{P}}(e^{\int_s^t (i\alpha-F_r)^*\, dW_r - \frac{1}{2}\int_s^t \|i\alpha-F_r\|^2\, dr}|\mathcal{F}_s)) \\ &\quad= e^{-\|\alpha\|^2(t-s)/2}\,\mathbb{E}_{\mathbb{P}}(\mathbb{E}_{\mathbb{P}}(\Lambda_T e^{i\beta Z}|\mathcal{F}_s)) = e^{-\|\alpha\|^2(t-s)/2}\,\mathbb{E}_{\mathbb{Q}}(e^{i\beta Z}).\end{aligned}$$

As the characteristic function factorizes into the characteristic function of Z and the characteristic function of an m-dimensional Gaussian random variable with mean zero and covariance $(t-s)I$, we find that $X_t - X_s$ is independent of Z and has the desired distribution. Hence we have proved the theorem for the case that F_t is bounded.

Let us now tackle the general case where F_t is not necessarily bounded. Define the processes $G_t^n = F_t I_{\|F_t\| \le n}$; then G_t^n is bounded for any n. Moreover,

$$\exp\left(\mathbb{E}_{\mathbb{P}}\left[\frac{1}{2}\int_0^T \|F_t\|^2\, dt\right]\right) \le \mathbb{E}_{\mathbb{P}}\left[\exp\left(\frac{1}{2}\int_0^T \|F_t\|^2\, dt\right)\right] < \infty$$

using Jensen's inequality and Novikov's condition, so we obtain

$$\mathbb{E}_{\mathbb{P}}\left[\int_0^T \|G_t^n - F_t\|^2\, dt\right] = \mathbb{E}_{\mathbb{P}}\left[\int_0^T \|F_t\|^2\, I_{\|F_t\|>n}\, dt\right] \xrightarrow{n\to\infty} 0$$

where we have used dominated convergence. In particular, we can choose a subsequence $m(n) \nearrow \infty$ so that the stochastic integral $I_t(G_\cdot^n)$ converges a.s. to $I_t(F_\cdot)$.

[7] *Abuse of notation alert:* here x^* means the transpose of the vector x, not the conjugate transpose; in particular, $\alpha^* = \alpha$ for all $\alpha \in \mathbb{C}$. Similarly, $\|x\|^2$ denotes $\sum_i (x_i)^2$. We have not defined complex Itô integrals, but you may set $\int_0^t (a_s + ib_s)\, dW_s \equiv \int_0^t a_s\, dW_s + i\int_0^t b_s\, dW_s$ when a_s and b_s are real-valued, i.e., the complex integral is the linear extension of the real integral. As this is the only place, other than in theorem 3.2.5, where we will use complex numbers, we will put up with our lousy notation just this once.

Let us continue with this subsequence, i.e., we define $F_t^n = G_t^{m(n)}$. For any n, define X_t^n, Λ_t^n, $\Lambda^n = \Lambda_T^n$ and $\mathbb{Q}^n$ by replacing F_t by F_t^n in their definitions. Then X_t^n is a Wiener process under $\mathbb{Q}^n$ for any $n < \infty$. In particular, for any n and $t > s$,

$$\mathbb{E}_{\mathbb{P}}(\Lambda^n\, e^{i\alpha^*(X_t^n - X_s^n)+i\beta Z}) = e^{-\|\alpha\|^2(t-s)/2}\,\mathbb{E}_{\mathbb{P}}(\Lambda^n\, e^{i\beta Z})$$

where Z is some $\mathcal{F}_s$-measurable random variable. We claim that[8]

$$\Lambda^n e^{i\alpha^*(X_t^n - X_s^n)+i\beta Z} \to \Lambda\, e^{i\alpha^*(X_t - X_s)+i\beta Z} \quad \text{and} \quad \Lambda^n e^{i\beta Z} \to \Lambda\, e^{i\beta Z} \quad \text{in } \mathcal{L}^1(\mathbb{P}).$$

Once this has been established, we are done: taking the limit as $n \to \infty$, we find that $\mathbb{E}_{\mathbb{Q}}(e^{i\alpha^*(X_t - X_s)+i\beta Z}) = e^{-\|\alpha\|^2(t-s)/2}\mathbb{E}_{\mathbb{Q}}(e^{i\beta Z})$ which is precisely what we want to show.

To proceed, let us first show that $\Lambda_n \to \Lambda$ in $\mathcal{L}^1(\mathbb{P})$. Note that $\Lambda_n \to \Lambda$ a.s., and using theorem 4.5.8 we find that $\mathbb{E}_{\mathbb{P}}(\Lambda_n) = \mathbb{E}_{\mathbb{P}}(\Lambda) = 1$. As $(\Lambda_n - \Lambda)^- \le \Lambda$ and $\Lambda_n \to \Lambda$ a.s., we find that $\mathbb{E}_{\mathbb{P}}((\Lambda_n - \Lambda)^-) \to 0$. Similarly, $\mathbb{E}_{\mathbb{P}}((\Lambda_n - \Lambda)^+) = \mathbb{E}_{\mathbb{P}}(\Lambda_n - \Lambda + (\Lambda_n - \Lambda)^-) \to 0$. Hence $\mathbb{E}_{\mathbb{P}}(|\Lambda_n - \Lambda|) = \mathbb{E}_{\mathbb{P}}((\Lambda_n - \Lambda)^+ + (\Lambda_n - \Lambda)^-) \to 0$, and we have determined that $\Lambda_n \to \Lambda$ in $\mathcal{L}^1(\mathbb{P})$. (The previous argument is also known as *Scheffé's lemma*.)

The remaining work is easy. Clearly $\Lambda^n e^{i\beta Z} \to \Lambda e^{i\beta Z}$ in $\mathcal{L}^1(\mathbb{P})$, as $e^{i\beta Z}$ is independent of n and has bounded real and imaginary parts. To deal with $\Lambda^n e^{i\alpha^*(X_t^n - X_s^n)+i\beta Z}$, note that its real and imaginary parts are of the form $\alpha_n\beta_n$ where α_n is bounded and converges a.s. to α, while β_n converges to β in $\mathcal{L}^1(\mathbb{P})$. But then the following expression converges to zero:

$$\mathbb{E}_{\mathbb{P}}(|\alpha_n\beta_n - \alpha\beta|) \le \mathbb{E}_{\mathbb{P}}(|\alpha_n - \alpha|\,|\beta|) + \mathbb{E}_{\mathbb{P}}(|\alpha_n|\,|\beta_n - \beta|) \le \mathbb{E}_{\mathbb{P}}(|\alpha_n - \alpha|\,|\beta|) + K\,\mathbb{E}_{\mathbb{P}}(|\beta_n - \beta|)$$

using dominated convergence for the first term on the right and $\mathcal{L}^1(\mathbb{P})$ convergence for the second. Hence $\Lambda^n e^{i\alpha^*(X_t^n - X_s^n)+i\beta Z} \to \Lambda\, e^{i\alpha^*(X_t - X_s)+i\beta Z}$ in $\mathcal{L}^1(\mathbb{P})$, and we are done. □

The technical lemmas are next.

Lemma 4.5.5. *Let M_t, $t \in [0,T]$ be a nonnegative local martingale. Then M_t is a supermartingale. In particular, if $\mathbb{E}(M_T) = \mathbb{E}(M_0)$, then M_t is a martingale.*

Proof. Let τ_n be a reducing sequence for M_t. Then using Fatou's lemma (note that $M_t \ge 0$), we obtain[9] for $T \ge t \ge s \ge 0$

$$\mathbb{E}(M_t|\mathcal{F}_s) = \mathbb{E}\left(\liminf_{n\to\infty} M_{t\wedge\tau_n}\,\middle|\,\mathcal{F}_s\right) \le \liminf_{n\to\infty} \mathbb{E}(M_{t\wedge\tau_n}|\mathcal{F}_s) = \liminf_{n\to\infty} M_{s\wedge\tau_n} = M_s.$$

In particular, M_t is a supermartingale. But a supermartingale which is not a martingale must have $\mathbb{P}(\mathbb{E}(M_t|\mathcal{F}_s) < M_s) > 0$ for some $t > s$. Hence if M_t is not a martingale, then $\mathbb{E}(M_T) \le \mathbb{E}(M_t) < \mathbb{E}(M_s) \le \mathbb{E}(M_0)$. But we have assumed that $\mathbb{E}(M_T) = \mathbb{E}(M_0)$. □

Lemma 4.5.6. *Let F_t be Itô integrable and let W_t be a Wiener process. Then*

$$\mathbb{E}\left[\exp\left(\frac{1}{2}\int_0^t F_s\,dW_s\right)\right] \le \sqrt{\mathbb{E}\left[\exp\left(\frac{1}{2}\int_0^t (F_s)^2\,ds\right)\right]}.$$

[8] By convergence in $\mathcal{L}^1(\mathbb{P})$, we mean that the real and imaginary parts converge in $\mathcal{L}^1(\mathbb{P})$.

[9] To be precise, we should first check that $M_t \in \mathcal{L}^1(\mathbb{P})$, otherwise the conditional expectation is not defined. But an identical application of Fatou's lemma shows that $\mathbb{E}(M_t) \le \mathbb{E}(M_0)$, so $M_t \in \mathcal{L}^1(\mathbb{P})$.

Proof. As $\exp(\frac{1}{2}\int_0^t F_s dW_s) = \exp(\frac{1}{2}\int_0^t F_s dW_s - \frac{1}{4}\int_0^t (F_s)^2 ds)\ \exp(\frac{1}{4}\int_0^t (F_s)^2 ds)$,

$$\mathbb{E}\left[\exp\left(\frac{1}{2}\int_0^t F_s\, dW_s\right)\right] \leq \sqrt{\mathbb{E}\left[\exp\left(\frac{1}{2}\int_0^t (F_s)^2\, ds\right)\right]\mathbb{E}\left[e^{\int_0^t F_s\, dW_s - \frac{1}{2}\int_0^t (F_s)^2\, ds}\right]}$$

using Hölder's inequality. But using Itô's rule, we find that

$$e^{\int_0^t F_s\, dW_s - \frac{1}{2}\int_0^t (F_s)^2\, ds} \equiv M_t = 1 + \int_0^t M_s\, F_s\, dW_s,$$

so evidently M_t is a local martingale. Hence M_t is a supermartingale by lemma 4.5.5, and this implies in particular $\mathbb{E}(M_t) \leq 1$. The result follows immediately. □

It remains to prove Novikov's theorem, i.e., that the Novikov condition implies that Λ_t is a martingale; this was key for the Girsanov theorem to hold. Evidently it suffices to prove that $\mathbb{E}_{\mathbb{P}}(\Lambda_T) = 1$ (lemma 4.5.5). To show this, let us introduce the following supporting lemma; the proof of Novikov's theorem then essentially amounts to reducing the Novikov condition to this lemma.

Lemma 4.5.7. *Let M_t be a nonnegative local martingale and let τ_n be a reducing sequence. If $\sup_n \|M_{T\wedge\tau_n}\|_p < \infty$ for some $p > 1$, then $\{M_t\}_{t\in[0,T]}$ is a martingale.*

Proof. We begin by writing

$$\mathbb{E}(|M_T - M_{T\wedge\tau_n}|) \leq \mathbb{E}(M_T - r\wedge M_T) + \mathbb{E}(|r\wedge M_T - r\wedge M_{T\wedge\tau_n}|) + \mathbb{E}(M_{T\wedge\tau_n} - r\wedge M_{T\wedge\tau_n}).$$

We claim that if we take the limit as $n\to\infty$ and as $r\to\infty$, in that order, then the right-hand side vanishes. This is clear for the first two terms, using dominated convergence and the fact that $\mathbb{E}(M_T) \leq \mathbb{E}(M_0) < \infty$ by the supermartingale property. To tackle the last term, note that

$$\mathbb{E}(M_{T\wedge\tau_n} - r\wedge M_{T\wedge\tau_n}) = \mathbb{E}(M_{T\wedge\tau_n} I_{M_{T\wedge\tau_n}>r}) - r\,\mathbb{P}(M_{T\wedge\tau_n} > r).$$

Using Chebyshev's inequality,

$$r\,\mathbb{P}(M_{T\wedge\tau_n} > r) \leq \frac{r\,\mathbb{E}((M_{T\wedge\tau_n})^p)}{r^p} \leq \frac{\sup_n \mathbb{E}((M_{T\wedge\tau_n})^p}{r^{p-1}},$$

so as $n\to\infty$ and $r\to\infty$ this term vanishes. Now let $0 < r \leq x$; then we have the trivial estimate $x \leq r^{1-p}x^p$ for $p > 1$. Hence

$$\mathbb{E}(M_{T\wedge\tau_n} I_{M_{T\wedge\tau_n}>r}) \leq r^{1-p}\,\mathbb{E}((M_{T\wedge\tau_n})^p I_{M_{T\wedge\tau_n}>r}) \leq r^{1-p}\sup_n \mathbb{E}((M_{T\wedge\tau_n})^p),$$

so this term also vanishes in the limit. Hence $\mathbb{E}(|M_T - M_{T\wedge\tau_n}|) \to 0$ as $n\to\infty$. But then $\mathbb{E}(M_T) = \lim_{n\to\infty}\mathbb{E}(M_{T\wedge\tau_n}) = \mathbb{E}(M_0)$, so M_t is a martingale by lemma 4.5.5. □

Theorem 4.5.8 (Novikov). *For Itô integrable F_t, define the local martingale*

$$\mathcal{E}_t(F_\cdot) = \exp\left[\int_0^t (F_s)^* dW_s - \frac{1}{2}\int_0^t \|F_s\|^2\, ds\right].$$

Suppose furthermore that the following condition is satisfied:

$$\mathbb{E}\left[\exp\left(\frac{1}{2}\int_0^T \|F_s\|^2\, ds\right)\right] = K < \infty.$$

Then $\{\mathcal{E}_t(F_\cdot)\}_{t\in[0,T]}$ is in fact a martingale.

Proof. We are going to show below that $\mathcal{E}_t(F.\sqrt{\alpha})$ is a martingale for all $0 < \alpha < 1$. Let us first argue that the result follows from this. By inspection, we easily find that

$$\mathcal{E}_t(F.\sqrt{\alpha}) = (\mathcal{E}_t(F.))^\alpha e^{\sqrt{\alpha}(1-\sqrt{\alpha})\int_0^t (F_s)^* dW_s}.$$

Applying Hölder's inequality with $p = \alpha^{-1}$ and $q = (1-\alpha)^{-1}$, we obtain

$$1 = \mathbb{E}(\mathcal{E}_T(F.\sqrt{\alpha})) \le (\mathbb{E}(\mathcal{E}_T(F.)))^\alpha (\mathbb{E}(e^{\sqrt{\alpha}(1+\sqrt{\alpha})^{-1}\int_0^T (F_s)^* dW_s}))^{1-\alpha}.$$

But as $x^{(1+\sqrt{\alpha})/2\sqrt{\alpha}}$ is a convex function for all $0 < \alpha < 1$, we find

$$1 \le (\mathbb{E}(\mathcal{E}_T(F.)))^\alpha (\mathbb{E}(e^{\frac{1}{2}\int_0^T (F_s)^* dW_s}))^{2\sqrt{\alpha}(1-\sqrt{\alpha})} \le (\mathbb{E}(\mathcal{E}_T(F.)))^\alpha K^{\sqrt{\alpha}(1-\sqrt{\alpha})},$$

where we have used lemma 4.5.6. Letting $\alpha \nearrow 1$, we find that $\mathbb{E}(\mathcal{E}_T(F.)) \ge 1$. But we already know that $\mathbb{E}(\mathcal{E}_T(F.)) \le 1$ by lemma 4.5.5, so the result follows.

It remains to show that $\mathcal{E}_t(F.\sqrt{\alpha})$ is a martingale for any $0 < \alpha < 1$. Fix α. As $\mathcal{E}_t(F.\sqrt{\alpha})$ is a local martingale, we can choose a localizing sequence τ_n. By lemma 4.5.7, it suffices to prove that $\sup_n \mathbb{E}((\mathcal{E}_{T\wedge\tau_n}(F.\sqrt{\alpha}))^u) < \infty$ for some $u > 1$. But exactly as above, we find

$$(\mathcal{E}_t(F.\sqrt{\alpha}))^u = (\mathcal{E}_t(F.))^{u\alpha} e^{u\sqrt{\alpha}(1-\sqrt{\alpha})\int_0^t (F_s)^* dW_s}.$$

Applying Hölder's inequality with $p = (u\alpha)^{-1}$ and $q = (1-u\alpha)^{-1}$, we obtain

$$\mathbb{E}((\mathcal{E}_{T\wedge\tau_n}(F.\sqrt{\alpha}))^u) \le (\mathbb{E}(\mathcal{E}_{T\wedge\tau_n}(F.)))^{u\alpha} (\mathbb{E}(e^{u\sqrt{\alpha}(1-\sqrt{\alpha})(1-u\alpha)^{-1}\int_0^{T\wedge\tau_n} (F_s)^* dW_s}))^{1-u\alpha}.$$

We should choose a suitable $u > 1$. Let us try

$$\frac{u\sqrt{\alpha}(1-\sqrt{\alpha})}{1-u\alpha} = \frac{1}{2} \qquad \Longrightarrow \qquad u = \frac{1}{2\sqrt{\alpha}(1-\sqrt{\alpha})+\alpha} = \frac{1}{(2-\sqrt{\alpha})\sqrt{\alpha}}.$$

But this actually works, because $0 < (2-x)x < 1$ for $0 < x < 1$, so $u > 1$ for any $0 < \alpha < 1$. Hence, choosing this particular u, we find that

$$\mathbb{E}((\mathcal{E}_{T\wedge\tau_n}(F.\sqrt{\alpha}))^u) \le (\mathbb{E}(e^{\frac{1}{2}\int_0^{T\wedge\tau_n} (F_s)^* dW_s}))^{1-u\alpha} \le K^{(1-u\alpha)/2},$$

where we have used that $\mathcal{E}_t(F.)$ is a supermartingale and that lemma 4.5.6 still holds when t is replaced by $t\wedge\tau_n$ (just replace F_s by $F_s I_{s<\tau_n}$, then apply lemma 4.5.6). Taking the supremum over n, we obtain what we set out to demonstrate. The proof is complete. □

4.6 The martingale representation theorem

We have one more fundamental result to take care of in this chapter. It is the martingale representation theorem—an altogether remarkable result. You know that the Itô integral of any integrand in $\mathcal{L}^2(\mu_T \times \mathbb{P})$ is a martingale which is in $\mathcal{L}^2(\mathbb{P})$ at every time t. In essence, the martingale representation theorem states precisely the converse: every martingale $\{M_t\}_{t\in[0,T]}$ with $M_T \in \mathcal{L}^2(\mathbb{P})$ can be represented as the Itô integral of a unique process in $\mathcal{L}^2(\mu_T \times \mathbb{P})$! Beside its fundamental interest, this idea plays an important role in mathematical finance; it also forms the basis for one approach to the filtering problem, though we will follow an alternative route in chapter 7.

We have to be a little bit careful by what we mean by a martingale: after all, we have not yet specified a filtration. For this particular result, we have to restrict

ourselves to a more limited setting than in the rest of the chapter. We will work on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which is defined a Wiener process W_t (we will take it to be one-dimensional, though you can easily extend the result to higher dimensions). The restriction will come in due to the fact that we only consider the natural filtration $\mathcal{F}_t^W = \sigma\{W_s : s \leq t\}$, unlike in the rest of the chapter where we could work with a larger filtration $\mathcal{F}_t$. The statement of the theorem is then as follows.

Theorem 4.6.1 (Martingale representation). *Let M_t be an $\mathcal{F}_t^W$-martingale such that $M_T \in \mathcal{L}^2(\mathbb{P})$. Then for a unique $\mathcal{F}_t^W$-adapted process $\{H_t\}_{t\in[0,T]}$ in $\mathcal{L}^2(\mu_T \times \mathbb{P})$*

$$M_t = M_0 + \int_0^t H_s \, dW_s \quad \text{a.s.} \quad \text{for all } t \in [0,T],$$

where the uniqueness of H_t is meant up to a $\mu_T \times \mathbb{P}$-null set.

Actually, the theorem is a trivial corollary of the following result.

Theorem 4.6.2 (Itô representation). *Let X be an $\mathcal{F}_T^W$-measurable random variable in $\mathcal{L}^2(\mathbb{P})$. Then for a unique $\mathcal{F}_t^W$-adapted process $\{H_t\}_{t\in[0,T]}$ in $\mathcal{L}^2(\mu_T \times \mathbb{P})$*

$$X = \mathbb{E}(X) + \int_0^T H_s \, dW_s \quad \text{a.s.},$$

where the uniqueness of H_t is meant up to a $\mu_T \times \mathbb{P}$-null set.

We will prove this theorem below. Let us first show, however, how the martingale representation theorem follows from the Itô representation theorem.

Proof of theorem 4.6.1. By theorem 4.6.2, we can write

$$M_T = \mathbb{E}(M_T) + \int_0^T H_s \, dW_s \quad \text{a.s.}$$

for a unique $\mathcal{F}_t^W$-adapted process $\{H_t\}_{t\in[0,T]}$ in $\mathcal{L}^2(\mu_T \times \mathbb{P})$. It remains to note that $\mathbb{E}(M_T) = M_0$ (as $\mathcal{F}_0^W$ is the trivial filtration, M_0 must be non-random) and that the martingale representation result follows from the Itô representation of M_T using $M_t = \mathbb{E}(M_T|\mathcal{F}_t^W)$. □

How should we go about proving the Itô representation theorem? We begin by proving an "approximate" version of the theorem: rather than showing that any $\mathcal{F}_T^W$-measurable random variable $X \in \mathcal{L}^2(\mathbb{P})$ can be represented as an Itô integral, we will show that any such X can be approximated arbitrarily well by an Itô integral, in the following sense: for arbitrarily small $\varepsilon > 0$, we can find an $\mathcal{F}_t^W$-adapted process H_t^ε in $\mathcal{L}^2(\mu_T \times \mathbb{P})$ such that $\|X - I_T(H_\cdot^\varepsilon)\|_2 < \varepsilon$. Once we have established this approximate version of the result, we are almost done: making the theorem exact rather than approximate is then simply a matter of taking limits.

The proof of the approximate Itô representation proceeds in two parts. First, we identify a class of random variables which can approximate any $\mathcal{F}_T^W$-measurable $X \in \mathcal{L}^2(\mathbb{P})$ arbitrarily well. Next, we show that any random variable in this class can be represented as the Itô integral of some $\mathcal{F}_t^W$-adapted process H_t in $\mathcal{L}^2(\mu_T \times \mathbb{P})$.

Lemma 4.6.3. *Introduce the following class of $\mathcal{F}_T^W$-measurable random variables:*

$$\mathcal{S} = \{f(W_{t_1}, \dots, W_{t_n}) : n < \infty,\ t_1, \dots, t_n \in [0,T],\ f \in C_0^\infty\}$$

(recall that C_0^∞ is the class of smooth functions with compact support). Then for any $\varepsilon > 0$ and $\mathcal{F}_T^W$-measurable $X \in \mathcal{L}^2(\mathbb{P})$, there is a $Y \in \mathcal{S}$ such that $\|X - Y\|_2 < \varepsilon$.

Proof. First, we claim that the statement holds if f is just assumed to be Borel-measurable rather than C_0^∞. To show this, introduce the filtration $\mathcal{G}_n = \sigma\{W_{k2^{-n}T} : k = 0, \dots, 2^n\}$, and note that $\mathcal{F}_T^W = \sigma\{\mathcal{G}_n : n = 1, 2, \dots\}$. Fix some $\mathcal{F}_T^W$-measurable $X \in \mathcal{L}^2(\mathbb{P})$, and define the sequence $X^n = \mathbb{E}(X|\mathcal{G}_n)$. Then $X^n \to X$ in $\mathcal{L}^2(\mathbb{P})$ by lemma 4.6.4 below. But $X^n = f(W_{2^{-n}T}, \dots, W_T)$ for some Borel function f (as it is $\mathcal{G}_n$-measurable), so X can be approximated arbitrarily closely by Borel functions of the Wiener process at a finite number of times. Note that we may also restrict ourselves to bounded Borel functions: after all, the sequence $X^n \wedge n$ of bounded random variables converges to X as well.

We now claim that any bounded Borel function f can be approximated arbitrarily well by functions $f^n \in C^\infty$. But this is well known: the approximations f^n can be found, e.g., by convolving f with a smooth function of compact support, and $f^n(W_{2^{-n}T}, \dots, W_T) \to f(W_{2^{-n}T}, \dots, W_T)$ in $\mathcal{L}^2(\mathbb{P})$ by dominated convergence. It remains to note that we can restrict ourselves to functions in C_0^∞, as we can always multiply f^n by g^n, where g^n is a sequence of $[0,1]$-valued functions with compact support such that $g^n \nearrow 1$ pointwise, and dominated convergence still gives the desired result. We are done. □

In the previous proof we used the following fundamental result.

Lemma 4.6.4 (Lévy's upward theorem). *Let $X \in \mathcal{L}^2(\mathbb{P})$ be $\mathcal{G}$-measurable, and let $\mathcal{G}_n$ be a filtration such that $\mathcal{G} = \sigma\{\mathcal{G}_n\}$. Then $\mathbb{E}(X|\mathcal{G}_n) \to X$ a.s. and in $\mathcal{L}^2(\mathbb{P})$.*

Proof. Let us write $X^n = \mathbb{E}(X|\mathcal{G}_n)$. Using Jensen's inequality is it easy to see that $\|X^n\|_2 \leq \|X\|_2 < \infty$, so $X^n \in \mathcal{L}^2$ for all n. In particular, as X^n is a martingale and $\sup_n \|X^n\|_1 < \infty$, it follow from the martingale convergence theorem that $X^n \to X^\infty$ a.s. We would like to prove that $X^\infty = X$ and that the convergence is in $\mathcal{L}^2(\mathbb{P})$ as well.

Let us first prove that $X \to X^\infty$ in $\mathcal{L}^2(\mathbb{P})$. Note that using the martingale property, $\mathbb{E}((X^n - X^m)^2) = \mathbb{E}((X^n)^2) - \mathbb{E}((X^m)^2)$ for $n \geq m$. But then $\mathbb{E}((X^n - X^m)^2) = \sum_{k=m}^{n-1} \mathbb{E}((X^{k+1} - X^k)^2)$. However, we know that the sum must converge as $n \to \infty$, as $\sup_n \|X^n\|_2 < \infty$. Hence X^n is Cauchy in $\mathcal{L}^2(\mathbb{P})$, so we have $X^n \to X^\infty$ in $\mathcal{L}^2(\mathbb{P})$ also.

It remains to prove $X^\infty = X$. Assume without loss of generality that $X \geq 0$ and that $\mathbb{E}(X) = 1$. Then $\mathbb{Q}(A) = \mathbb{P}(XI_A)$ is another probability measure, and we claim that $\mathbb{Q}(A) = \mathbb{P}(X^\infty I_A)$ as well for all $A \in \mathcal{G}$. This would establish the claim by the uniqueness of the Radon-Nikodym derivative. But note that $\bigcup_n \mathcal{G}_n$ is a π-system that generates $\mathcal{G}$, so it suffices to check the condition for A in this π-system. But for any such A we have $\mathbb{E}(XI_A) = \mathbb{E}(X^n I_A)$ for sufficiently large n, and as $X^n \to X^\infty$ in $\mathcal{L}^2(\mathbb{P})$ we find that $\mathbb{E}(X^\infty I_A) = \lim_{n\to\infty} \mathbb{E}(X^n I_A) = \mathbb{E}(XI_A)$. The proof is complete. □

We now come to the point of this exercise.

Lemma 4.6.5 (Approximate Itô representation). *For any $Y \in \mathcal{S}$, there is an $\mathcal{F}_t^W$-adapted process H_t in $\mathcal{L}^2(\mu_T \times \mathbb{P})$ such that*

$$Y = \mathbb{E}(Y) + \int_0^T H_s\, dW_s.$$

In particular, this implies that any $\mathcal{F}_T^W$-measurable random variable $X \in \mathcal{L}^2(\mathbb{P})$ can be approximated arbitrarily closely in $\mathcal{L}^2(\mathbb{P})$ by an Itô integral.

Proof. Let us first consider the simplest case $Y = f(W_t)$, where $f \in C_0^\infty$ and $t \in [0,T]$. Note that for any function $g(s,x)$ which is sufficiently smooth, we have by Itô's rule

$$g(t,W_t) = g(0,0) + \int_0^t \left[\frac{\partial g}{\partial s} + \frac{1}{2}\frac{\partial^2 g}{\partial x^2}\right](s,W_s)\,ds + \int_0^t \frac{\partial g}{\partial x}(s,W_s)\,dW_s.$$

The trick is to choose the function $g(s,x)$ precisely so that

$$\frac{\partial g}{\partial s} + \frac{1}{2}\frac{\partial^2 g}{\partial x^2} = 0, \qquad g(t,x) = f(x).$$

But such a function exists and is even sufficiently smooth: explicitly,

$$g(s,x) = \frac{1}{\sqrt{2\pi(t-s)}} \int_{-\infty}^{\infty} f(y)\, e^{-(x-y)^2/2(t-s)}\,dy, \qquad s < t.$$

Hence the integrand $H_s = (\partial g/\partial x)(s,W_s)$ gets the job done, and is clearly $\mathcal{F}_t^W$-adapted. It is also in $\mathcal{L}^2(\mu_T \times \mathbb{P})$ as f is of compact support, so $\partial g/\partial x$ is of compact support as well (and in particular bounded). Hence in the very simplest case, we are done.

Let us now consider the next most difficult case: $Y = f(W_r, W_t)$ with $r < t$. Introduce $g(s,x,z)$, and apply Itô's rule to $g(s, W_{s\wedge r}, W_s)$. We find that

$$g(t,W_r,W_t) = g(r,W_r,W_r) + \int_r^t \left[\frac{\partial g}{\partial s} + \frac{1}{2}\frac{\partial^2 g}{\partial z^2}\right](s,W_r,W_s)\,ds + \int_r^t \frac{\partial g}{\partial z}(s,W_r,W_s)\,dW_s.$$

But for some function $g'(s,x)$, applying Itô's rule to $g'(s,W_s)$ gives

$$g'(r,W_r) = g'(0,0) + \int_0^r \left[\frac{\partial g'}{\partial s} + \frac{1}{2}\frac{\partial^2 g'}{\partial x^2}\right](s,W_s)\,ds + \int_0^r \frac{\partial g'}{\partial x}(s,W_s)\,dW_s.$$

Thus evidently we wish to solve the partial differential equations

$$\frac{\partial g}{\partial s} + \frac{1}{2}\frac{\partial^2 g}{\partial z^2} = 0, \quad g(t,x,z) = f(x,z), \qquad \frac{\partial g'}{\partial s} + \frac{1}{2}\frac{\partial^2 g'}{\partial x^2} = 0, \quad g'(r,x) = g(r,x,x).$$

This can be done exactly as before, and we set

$$H_t = \frac{\partial g'}{\partial x}(s,W_s) \quad \text{for } 0 \le s \le r, \qquad H_t = \frac{\partial g}{\partial z}(s,W_r,W_s) \quad \text{for } r \le s \le t.$$

Proceeding in the same manner, we find by induction that the result holds for any $Y \in \mathcal{S}$. □

We can now complete the proof of the Itô representation theorem.

Proof of theorem 4.6.2. Let X be $\mathcal{F}_T^W$-measurable and in $\mathcal{L}^2(\mathbb{P})$, and choose a sequence $X^n \in \mathcal{S}$ such that $\|X - X^n\|_2 \to 0$. We may assume without loss of generality that $\mathbb{E}(X) = 0$ and $\mathbb{E}(X^n) = 0$ for all n. By the previous lemma, every X^n can be represented as the Itô integral of an $\mathcal{F}_t^W$-adapted process H_t^n in $\mathcal{L}^2(\mu_T \times \mathbb{P})$. In particular, we find that

$$\mathbb{E}((X^n - X^m)^2) = \mathbb{E}\left[\int_0^T (H_s^n - H_s^m)^2\,ds\right].$$

But as $X^n \to X$ in $\mathcal{L}^2(\mathbb{P})$, it must be the case that $\mathbb{E}((X^n - X^m)^2) \to 0$ as $m, n \to \infty$. Hence $H_\cdot^n$ is a Cauchy sequence in $\mathcal{L}^2(\mu_T \times \mathbb{P})$, and so has a limit process $H_\cdot$ in $\mathcal{L}^2(\mu_T \times \mathbb{P})$. Moreover, as each H_t^n is adapted, so is H_t. Using the Itô isometry, it is easy to verify that

$$X = \int_0^T H_s \, dW_s \qquad a.s.$$

Hence we conclude that X is representable as an Itô integral, as promised.

It only remains to prove uniqueness. Suppose that H_s and H'_s both fit the bill. Then

$$\mathbb{E}\left[\int_0^T (H_s - H'_s)^2 \, ds\right] = 0$$

by the Itô isometry. But this can only be the case if $H_\cdot = H'_\cdot$ $\mu_T \times \mathbb{P}$-a.s. □

4.7 Further reading

As a general introductory textbook on stochastic calculus and its applications, the widely used book by Øksendal [Øks03] is an excellent, very readable introduction and is highly recommended. A slightly older but still excellent introductory text is the book by Arnold [Arn74] (now unfortunately out of print).

The discussion in the first section is inspired to some extent by a brief discussion of a similar nature in the book by Bichteler [Bic02]. Protter [Pro04] also has a brief discussion on the matter, but from a slightly different point of view.

There are many excellent advanced textbooks on stochastic calculus; however, most of these start directly with the general integration theory, where one can integrate against almost any martingale rather than just a Wiener process. See, for example, Rogers and Williams [RW00b], Karatzas and Shreve [KS91], Elliott [Ell82], Revuz and Yor [RY99], or Dellacherie and Meyer [DM82]. Perhaps the most natural approach to the general theory can be found in Protter [Pro04], who simply defines stochastic integrals against anything that makes sense, and then proceeds to show which processes indeed do. Bichteler [Bic02] follows a similar route.

The theory for the Wiener process was historically first [Itô44], and is still the most widely used (though the generalization to martingales began not so long after that—like many things in probability theory, this goes all the way back to [Doo53]). It is good to understand the "classical" Itô integral, which we have discussed in this chapter, before moving on to more advanced theory. Books which treat stochastic integration with respect to the Wiener process in significant detail are Friedman [Fri75] and Liptser and Shiryaev [LS01a]. The notion of localization, which is very fundamental in general stochastic integration, does not play a large role in those references; a nice discussion of localization in the Wiener process setting can be found in [Ste01].

The discussion of Girsanov's theorem is inspired by Friedman [Fri75]. In the modern stochastic integration theory, the Girsanov theorem has a much wider scope and indeed characterizes almost any change of measure; see, e.g., [Pro04].

Finally, our discussion of the martingale representation theorem follows Øksendal [Øks03], using an approach that is originally due to Davis [Dav80].

CHAPTER 5

Stochastic Differential Equations

Now that we have Itô integrals, we can introduce stochastic differential equations—one of the major reasons to set up the Itô theory in the first place. After dealing with issues of existence and uniqueness, we will exhibit an important property of stochastic differential equations: the solutions of such equations obey the Markov property. A particular consequence is the connection with the classic PDE methods for studying diffusions, the Kolmogorov forward (Fokker-Planck) and backward equations.

We would like to think of stochastic differential equations (SDE) as ordinary differential equations (ODE) driven by white noise. Unfortunately, this connection is not entirely clear; after all, we have only justified the connection between the Itô integral and white noise in the case of non-random integrands (interpreted as test functions). We will show that if we take a sequence of ODEs, driven by approximations to white noise, then these do indeed limit to an SDE—though not entirely the expected one. This issue is particularly important in the stochastic modelling of physical systems. A related issue is the simulation of SDE on a computer, which we will briefly discuss.

Finally, the chapter concludes with a brief discussion of some more advanced topics in stochastic differential equations.

5.1 Stochastic differential equations: existence and uniqueness

One often encounters stochastic differential equations written in the form

$$dX_t = b(t, X_t)\,dt + \sigma(t, X_t)\,dW_t, \qquad X_0 = x,$$

which looks almost like an ordinary differential equation. However, as usual, the "Itô differentials" are not sensible mathematical objects in themselves; rather, we should

see this expression as suggestive notation for the Itô process

$$X_t = x + \int_0^t b(s, X_s)\, ds + \int_0^t \sigma(s, X_s)\, dW_s.$$

If there exists a stochastic process X_t that satisfies this equation, we say that it *solves* the stochastic differential equation. The main goal of this section is to find conditions on the coefficients b an σ that guarantee the existence and uniqueness of solutions.

Example 5.1.1 (Linear SDE). Let W_t be an m-dimensional Wiener process, let A be an $n \times n$ matrix and let B be an $n \times m$ matrix. Then the n-dimensional equation

$$dX_t = AX_t\, dt + B\, dW_t, \qquad X_0 = x,$$

is called a *linear stochastic differential equation*. Such equations always have a solution; in fact, the solution can be given explicitly by

$$X_t = e^{At}x + \int_0^t e^{A(t-s)} B\, dW_s,$$

as you can verify directly by applying Itô's rule. The solution is also unique. To see this, let Y_t be another solution with the same initial condition. Then

$$X_t - Y_t = \int_0^t A(X_s - Y_s)\, ds \quad \Longrightarrow \quad \frac{d}{dt}(X_t - Y_t) = A(X_t - Y_t), \quad X_0 - Y_0 = 0,$$

and it is a standard fact that the unique solution of this equation is $X_t - Y_t = 0$.

For nonlinear b and σ one can rarely write down the solution in explicit form, so we have to resort to a less explicit proof of existence. The same problem appears in the theory of ordinary differential equations where it is often resolved by imposing Lipschitz conditions, whereupon existence can be proved by Picard iteration (see, e.g., [Apo69, theorem 7.19]). It turns out that this technique works almost identically in the current setting. Let us work out the details. Recall what it means to be Lipschitz:

Definition 5.1.2. A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is called *Lipschitz continuous* (or just Lipschitz) if there exists a constant $K < \infty$ such that $\|f(x) - f(y)\| \le K\|x - y\|$ for all $x, y \in \mathbb{R}^n$. A function $g : S \times \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz *uniformly in* S if $\|g(s, x) - g(s, y)\| \le K\|x - y\|$ for a constant $K < \infty$ which does not depend on s.

We work in the following setting, where we restrict ourselves to a finite time horizon $[0, T]$ for simplicity. Consider a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\in[0,T]}, \mathbb{P})$ on which is defined an m-dimensional $\mathcal{F}_t$-Wiener process W_t. We now choose X_0 to be an $\mathcal{F}_0$-measurable n-dimensional random variable (it is often chosen to be non-random, but this is not necessary), and we seek a solution to the equation

$$X_t = X_0 + \int_0^t b(s, X_s)\, ds + \int_0^t \sigma(s, X_s)\, dW_s.$$

Here $b : [0, T] \times \mathbb{R}^n \to \mathbb{R}^n$ and $\sigma : [0, T] \times \mathbb{R}^n \to \mathbb{R}^{n\times m}$ are at least measurable.

Theorem 5.1.3 (Existence). *Suppose that*

1. $X_0 \in \mathcal{L}^2(\mathbb{P})$*; and*
2. b, σ *are Lipschitz continuous uniformly on* $[0,T]$*; and*
3. $\|b(t,0)\|$ *and* $\|\sigma(t,0)\|$ *are bounded on* $t \in [0,T]$.

Then there exists a solution X_t *to the associated stochastic differential equation, and moreover for this solution* X_t*,* $b(t,X_t)$*, and* $\sigma(t,X_t)$ *are in* $\mathcal{L}^2(\mu_T \times \mathbb{P})$.

Proof. For any $\mathcal{F}_t$-adapted $Y_\cdot \in \mathcal{L}^2(\mu_T \times \mathbb{P})$, introduce the following (nonlinear) map:

$$(\mathfrak{P}(Y_\cdot))_t = X_0 + \int_0^t b(s, Y_s)\, ds + \int_0^t \sigma(s, Y_s)\, dW_s.$$

We claim that under the conditions which we have imposed, $\mathfrak{P}(Y_\cdot)$ is again an $\mathcal{F}_t$-adapted process in $\mathcal{L}^2(\mu_T \times \mathbb{P})$ (we will show this shortly). Our goal is to find a *fixed point* of the operator $\mathfrak{P}$: i.e., we wish to find an $\mathcal{F}_t$-adapted process $X_\cdot \in \mathcal{L}^2(\mu_T \times \mathbb{P})$ such that $\mathfrak{P}(X_\cdot) = X_\cdot$. Such an $X_\cdot$ is then, by definition, a solution of our stochastic differential equation.

We begin by showing that $\mathfrak{P}$ does indeed map to an $\mathcal{F}_t$-adapted process in $\mathcal{L}^2(\mu_T \times \mathbb{P})$. Note that $\|b(t,x)\| \le \|b(t,x) - b(t,0)\| + \|b(t,0)\| \le K\|x\| + K' \le C(1 + \|x\|)$ where $K, K', C < \infty$ are constants that do not depend on t. We say that b satisfies a *linear growth condition*. Clearly the same argument holds for σ, and to make our notation lighter we will choose our constant C such that $\|\sigma(t,x)\| \le C(1 + \|x\|)$ as well. We can now estimate

$$\|\mathfrak{P}(Y_\cdot)\|_{2,\mu_T\times\mathbb{P}} \le \|X_0\|_{2,\mu_T\times\mathbb{P}} + \left\|\int_0^\cdot b(s,Y_s)\, ds\right\|_{2,\mu_T\times\mathbb{P}} + \left\|\int_0^\cdot \sigma(s,Y_s)\, dW_s\right\|_{2,\mu_T\times\mathbb{P}}.$$

The first term gives $\|X_0\|_{2,\mu_T\times\mathbb{P}} = \sqrt{T}\,\|X_0\|_{2,\mathbb{P}} < \infty$ by assumption. Next,

$$\left\|\int_0^\cdot b(s,Y_s)\, ds\right\|^2_{2,\mu_T\times\mathbb{P}} \le T^2\,\|b(\cdot,Y_\cdot)\|^2_{2,\mu_T\times\mathbb{P}} \le T^2C^2\,\|(1+\|Y_\cdot\|)\|^2_{2,\mu_T\times\mathbb{P}} < \infty,$$

where we have used $(t^{-1}\int_0^t a_s\, ds)^2 \le t^{-1}\int_0^t a_s^2\, ds$ (Jensen's inequality), the linear growth condition, and $Y_\cdot \in \mathcal{L}^2(\mu_T \times \mathbb{P})$. Finally, let us estimate the stochastic integral term:

$$\left\|\int_0^\cdot \sigma(s,Y_s)\, dW_s\right\|^2_{2,\mu_T\times\mathbb{P}} \le T\,\|\sigma(\cdot,Y_\cdot)\|^2_{2,\mu_T\times\mathbb{P}} \le TC^2\,\|(1+\|Y_\cdot\|)\|^2_{2,\mu_T\times\mathbb{P}} < \infty,$$

where we have used the Itô isometry. Hence $\|\mathfrak{P}(Y_\cdot)\|_{2,\mu_T\times\mathbb{P}} < \infty$ for $\mathcal{F}_t$-adapted $Y_\cdot \in \mathcal{L}^2(\mu_T \times \mathbb{P})$, and clearly $\mathfrak{P}(Y_\cdot)$ is $\mathcal{F}_t$-adapted, so the claim is established.

Our next claim is that $\mathfrak{P}$ is a *continuous* map; i.e., we claim that if $\|Y_\cdot^n - Y_\cdot\|_{2,\mu_T\times\mathbb{P}} \to 0$, then $\|\mathfrak{P}(Y_\cdot^n) - \mathfrak{P}(Y_\cdot)\|_{2,\mu_T\times\mathbb{P}} \to 0$ as well. But proceeding exactly as before, we find that

$$\|\mathfrak{P}(Y_\cdot^n) - \mathfrak{P}(Y_\cdot)\|_{2,\mu_T\times\mathbb{P}} \le T\,\|b(\cdot,Y_\cdot^n) - b(\cdot,Y_\cdot)\|_{2,\mu_T\times\mathbb{P}} + \sqrt{T}\,\|\sigma(\cdot,Y_\cdot^n) - \sigma(\cdot,Y_\cdot)\|_{2,\mu_T\times\mathbb{P}}.$$

In particular, using the Lipschitz condition, we find that

$$\|\mathfrak{P}(Y_\cdot^n) - \mathfrak{P}(Y_\cdot)\|_{2,\mu_T\times\mathbb{P}} \le K\sqrt{T}\,(\sqrt{T}+1)\,\|Y_\cdot^n - Y_\cdot\|_{2,\mu_T\times\mathbb{P}},$$

where K is a Lipschitz constant for both b and σ. This establishes the claim.

With these preliminary issues out of the way, we now come to the heart of the proof. Starting from an arbitrary $\mathcal{F}_t$-adapted process Y_t^0 in $\mathcal{L}^2(\mu_T \times \mathbb{P})$, consider the sequence $Y_\cdot^1 = \mathfrak{P}(Y_\cdot^0)$, $Y_\cdot^2 = \mathfrak{P}(Y_\cdot^1) = \mathfrak{P}^2(Y_\cdot^0)$, etc. This is called *Picard iteration*. We will show below that $Y_\cdot^n$ is a Cauchy sequence in $\mathcal{L}^2(\mu_T \times \mathbb{P})$; hence it converges to some $\mathcal{F}_t$-adapted process Y_t in $\mathcal{L}^2(\mu_T \times \mathbb{P})$. But then $Y_\cdot$ is necessarily a fixed point of $\mathfrak{P}$: after all, $\mathfrak{P}(Y_\cdot^n) \to \mathfrak{P}(Y_\cdot)$ by the continuity of $\mathfrak{P}$, whereas $\mathfrak{P}(Y_\cdot^n) = Y_\cdot^{n+1} \to Y_\cdot$. Thus $\mathfrak{P}(Y_\cdot) = Y_\cdot$, and we have found a solution of our stochastic differential equation with the desired properties.

It only remains to show that $Y_\cdot^n$ is a Cauchy sequence in $\mathcal{L}^2(\mu_T \times \mathbb{P})$. This follows from a slightly refined version of the argument that we used to prove continuity of $\mathfrak{P}$. Note that

$$\|(\mathfrak{P}(Z_\cdot))_t - (\mathfrak{P}(Y_\cdot))_t\|_{2,\mathbb{P}} \le \sqrt{t}\,\|b(\cdot, Z_\cdot) - b(\cdot, Y_\cdot)\|_{2,\mu_t\times\mathbb{P}} + \|\sigma(\cdot, Z_\cdot) - \sigma(\cdot, Y_\cdot)\|_{2,\mu_t\times\mathbb{P}},$$

which follows exactly as above. In particular, using the Lipschitz property, we find

$$\|(\mathfrak{P}(Z_\cdot))_t - (\mathfrak{P}(Y_\cdot))_t\|_{2,\mathbb{P}} \le K(\sqrt{T}+1)\,\|Z_\cdot - Y_\cdot\|_{2,\mu_t\times\mathbb{P}}.$$

Set $L = K(\sqrt{T}+1)$. Iterating this bound, we obtain

$$\begin{aligned}
\|\mathfrak{P}^n(Z_\cdot) - \mathfrak{P}^n(Y_\cdot)\|^2_{2,\mu_T\times\mathbb{P}} &= \\
\int_0^T \|(\mathfrak{P}^n(Z_\cdot))_t - (\mathfrak{P}^n(Y_\cdot))_t\|^2_{2,\mathbb{P}}\,dt &\le L^2 \int_0^T \|\mathfrak{P}^{n-1}(Z_\cdot) - \mathfrak{P}^{n-1}(Y_\cdot)\|^2_{2,\mu_{t_1}\times\mathbb{P}}\,dt_1 \\
&\le \cdots \le L^{2n} \int_0^T \int_0^{t_1} \cdots \int_0^{t_{n-1}} \|Z_\cdot - Y_\cdot\|^2_{2,\mu_{t_n}\times\mathbb{P}}\,dt_n \cdots dt_1 \\
&\le \frac{L^{2n}T^n}{n!}\,\|Z_\cdot - Y_\cdot\|^2_{2,\mu_T\times\mathbb{P}}.
\end{aligned}$$

In particular, this implies that

$$\sum_{n=0}^\infty \|\mathfrak{P}^{n+1}(Y_\cdot^0) - \mathfrak{P}^n(Y_\cdot^0)\|_{2,\mu_T\times\mathbb{P}} \le \|\mathfrak{P}(Y_\cdot^0) - Y_\cdot^0\|_{2,\mu_T\times\mathbb{P}} \sum_{n=0}^\infty \sqrt{\frac{L^{2n}T^n}{n!}} < \infty,$$

which establishes that $\mathfrak{P}^n(Y_\cdot^0)$ is a Cauchy sequence in $\mathcal{L}^2(\mu_T \times \mathbb{P})$. We are done. □

Remark 5.1.4. The condition $X_0 \in \mathcal{L}^2(\mathbb{P})$ can be relaxed through localization, see [GS96, theorem VIII.3.1]; we then have a solution of the SDE for any initial condition, but we need no longer have X_t, $b(t, X_t)$, and $\sigma(t, X_t)$ in $\mathcal{L}^2(\mu_T \times \mathbb{P})$. More interesting, perhaps, is that if $X_0 \in \mathcal{L}^p(\mathbb{P})$ ($p \ge 2$), then we can prove with a little more work that X_t, $b(t, X_t)$, and $\sigma(t, X_t)$ will actually be in $\mathcal{L}^p(\mu_T \times \mathbb{P})$ (see [LS01a, sec. 4.4] or [Arn74, theorem 7.1.2]). Hence the integrability of the initial condition really determines the integrability of the solution in the Lipschitz setting.

It remains to prove uniqueness of the solution found in theorem 5.1.3.

Theorem 5.1.5 (Uniqueness). *The solution of theorem 5.1.3 is unique $\mathbb{P}$-a.s.*

Proof. Let $X_\cdot$ be the solution of theorem 5.1.3, and let $Y_\cdot$ be any other solution. It suffices to show that $X_\cdot = Y_\cdot$ $\mu_T \times \mathbb{P}$-a.s.; after all, both X_t and Y_t must have continuous sample paths, so $X_\cdot = Y_\cdot$ $\mu_T \times \mathbb{P}$-a.s. implies that they are $\mathbb{P}$-a.s. indistinguishable (lemma 2.4.6).

Let us first suppose that $Y_\cdot \in \mathcal{L}^2(\mu_T \times \mathbb{P})$ as well; then $\mathfrak{P}^n(Y_\cdot) = Y_\cdot$ and $\mathfrak{P}^n(X_\cdot) = X_\cdot$. Using the estimate in the proof of theorem 5.1.3, we find that

$$\|Y_\cdot - X_\cdot\|^2_{2,\mu_T\times\mathbb{P}} = \|\mathfrak{P}^n(Y_\cdot) - \mathfrak{P}^n(X_\cdot)\|^2_{2,\mu_T\times\mathbb{P}} \le \frac{L^{2n}T^n}{n!}\,\|Y_\cdot - X_\cdot\|^2_{2,\mu_T\times\mathbb{P}}.$$

Letting $n \to \infty$, we find that $\|Y_\cdot - X_\cdot\|_{2,\mu_T \times \mathbb{P}} = 0$, so $X_\cdot = Y_\cdot$ $\mu_T \times \mathbb{P}$-a.s.

We now claim that any solution Y_t with $Y_0 = X_0 \in \mathcal{L}^2(\mathbb{P})$ must necessarily be an element of $\mathcal{L}^2(\mu_T \times \mathbb{P})$; once this is established, the proof is complete. Let us write, using Itô's rule,

$$\|Y_t\|^2 = \|X_0\|^2 + \int_0^t (2(Y_s)^* b(s, Y_s) + \|\sigma(s, Y_s)\|^2)\, ds + \int_0^t 2(Y_s)^* \sigma(s, Y_s)\, dW_s.$$

Now let $\tau_n = \inf\{t : \|Y_t\| \geq n\}$, and note that this sequence of stopping times is a localizing sequence for the stochastic integral; in particular,

$$\mathbb{E}(\|Y_{t\wedge\tau_n}\|^2) = \mathbb{E}(\|X_0\|^2) + \mathbb{E}\left[\int_0^{t\wedge\tau_n} (2(Y_s)^* b(s, Y_s) + \|\sigma(s, Y_s)\|^2)\, ds\right].$$

Using the linear growth condition, we can now estimate

$$\mathbb{E}(\|Y_{t\wedge\tau_n}\|^2) \leq \mathbb{E}(\|X_0\|^2) + \mathbb{E}\left[\int_0^{t\wedge\tau_n} (2C\|Y_s\|(1 + \|Y_s\|) + C^2(1 + \|Y_s\|)^2)\, ds\right].$$

Using Fatou's lemma on the left and monotone convergence on the right to let $n \to \infty$, applying Tonelli's theorem, and using the simple estimate $(a + b)^2 \leq 2(a^2 + b^2)$, we obtain

$$\mathbb{E}(1 + \|Y_t\|^2) \leq \mathbb{E}(1 + \|X_0\|^2) + 2C(2 + C) \int_0^t \mathbb{E}(1 + \|Y_s\|^2)\, ds.$$

But then we find that $\mathbb{E}(1+\|Y_t\|^2) \leq \mathbb{E}(1+\|X_0\|^2)\, e^{2C(2+C)t}$ using Gronwall's lemma, from which the claim follows easily. Hence the proof is complete. □

5.2 The Markov property and Kolmogorov's equations

One of the most important properties of stochastic differential equations is that their solutions satisfy the *Markov property*. This means that a large class of Markov processes with continuous sample paths—these are important both from a fundamental and from an applied perspective—can be obtained as the solution of an appropriate SDE. Conversely, this means that methods from the theory of Markov processes can be used to study the properties of stochastic differential equations; in particular, the Komogorov equations (the forward, or Fokker-Planck, equation, and the backward equation) can be obtained in the SDE setting, and can be used to express expectations of functions of an SDE in terms of certain non-random PDEs.

Remark 5.2.1. The intricate theory of Markov processes in continuous time, like martingale theory, can easily fill an entire course on its own. It has deep connections with semigroup theory on the one hand, and probabilistic theory (at the level of sample paths) on the other hand. A development at this level would be way beyond the scope of these notes. We will content ourselves with proving the Markov property, and then developing the Kolmogorov equations in the simplest possible way (without invoking any theorems from the theory of Markov processes). If you are interested in the bigger picture, you might want to consult some of the references in section 5.7.

Let us begin by proving the Markov property.

Theorem 5.2.2 (Markov property). *Suppose that the conditions of theorem 5.1.3 hold. Then the unique solution X_t of the corresponding SDE is an $\mathcal{F}_t$-Markov process.*

Proof. Let us begin by rewriting the SDE in the following form:

$$X_t = X_s + \int_s^t b(r, X_r)\, dr + \int_s^t \sigma(r, X_r)\, dW_r,$$

which follows easily by calculating $X_t - X_s$. As X_s is $\mathcal{F}_s$-measurable and $\{W_{r+s} - W_s : r \geq 0\}$ is a Wiener process independent of $\mathcal{F}_s$, we can identically write this equation as

$$Y_{t-s} = Y_0 + \int_0^{t-s} \tilde{b}(r, Y_r)\, dr + \int_0^{t-s} \tilde{\sigma}(r, Y_r)\, d\tilde{W}_r,$$

where $Y_r = X_{r+s}$, $\tilde{b}(r,x) = b(r+s,x)$, $\tilde{\sigma}(r,x) = \sigma(r+s,x)$, and $\tilde{W}_r = W_{r+s} - W_s$. But this equation for Y_t is again an SDE that satisfies the conditions of theorems 5.1.3 and 5.1.5 in the interval $r \in [0, T-s]$, and in particular, it follows that Y_r is $\sigma\{Y_0, \tilde{W}_s : s \leq r\}$-measurable. Identically, we find that X_t is $\sigma\{X_s, W_r - W_s : r \in [s,t]\}$-measurable, and can hence be written as a measurable functional $X_t = F(X_s, W_{\cdot+s} - W_s)$. Now using Fubini's theorem exactly as in lemma 3.1.9, we find that $\mathbb{E}(g(X_t)|\mathcal{F}_s) = \mathbb{E}(g(F(x, W_{\cdot+s} - W_s)))|_{x=X_s}$ for any bounded measurable function g, so in particular $\mathbb{E}(g(X_t)|\mathcal{F}_s) = \mathbb{E}(g(X_t)|X_s)$ by the tower property of the conditional expectation. But this is the Markov property, so we are done. □

Remark 5.2.3 (Strong Markov property). The solutions of Lipschitz stochastic differential equations, and in particular the Wiener process itself, actually satisfy a much stronger variant of the Markov property. Let τ be an a.s. finite stopping time; then it turns out that $\mathbb{E}(g(X_{\tau+r})|\mathcal{F}_\tau) = \mathbb{E}(g(X_{\tau+r})|X_\tau)$. This is called the *strong Markov property*, which extends the Markov property even to random times. This fact is often very useful, but we will not prove it here; see, e.g., [Fri75, theorem 5.3.4].

The Markov property implies that for any bounded and measurable f, we have $\mathbb{E}(f(X_t)|\mathcal{F}_s) = g_{t,s}(X_s)$ for some (non-random) measurable function $g_{t,s}$. For the rest of this section, let us assume for simplicity that $b(t,x)$ and $\sigma(t,x)$ are independent of t (this is not essential, but will make the notation a little lighter); then you can read off from the previous proof that in fact $\mathbb{E}(f(X_t)|\mathcal{F}_s) = g_{t-s}(X_s)$ for some function g_{t-s}. We say that the Markov process is *time-homogeneous* in this case.

Rather than studying the random process X_t, we can now study how the non-random function g_t varies with t. This is a standard idea in the theory of Markov processes. Note that if $\mathbb{E}(f(X_t)|\mathcal{F}_s) = g_{t-s}(X_s)$ and $\mathbb{E}(f'(X_t)|\mathcal{F}_s) = g'_{t-s}(X_s)$, then $\mathbb{E}(\alpha f(X_t) + \beta f'(X_t)|\mathcal{F}_s) = \alpha g_{t-s}(X_s) + \beta g'_{t-s}(X_s)$; i.e., the map $P_t : f \mapsto g_t$ is linear. Moreover, using the tower property of the conditional expectation, you can easily convince yourself that $P_t g_s = g_{t+s}$. Hence $P_t P_s = P_{t+s}$, so the family $\{P_t\}$ forms a *semigroup*. This suggests[1] that we can write something like

$$\frac{d}{dt} P_t f = \mathscr{L} P_t f, \qquad P_0 f = f,$$

where $\mathscr{L}$ is a suitable linear operator. If such an equation holds for a sufficiently large class of functions f, then $\mathscr{L}$ is called the *infinitesimal generator* of the semigroup P_t.

[1] Think of a finite-dimensional semigroup $P_t x = e^{At} x$, where A is a square matrix and x is a vector.

Making these ideas mathematically sound is well beyond our scope; but let us show that under certain conditions, we can indeed obtain an equation of this form for $P_t f$.

Proposition 5.2.4 (Kolmogorov backward equation). *For $g \in C^2$, define*

$$\mathscr{L}g(x) = \sum_{i=1}^{n} b^i(x)\,\frac{\partial g}{\partial x^i}(x) + \frac{1}{2}\sum_{i,j=1}^{n}\sum_{k=1}^{m}\sigma^{ik}(x)\sigma^{jk}(x)\frac{\partial^2 g}{\partial x^i\,\partial x^j}(x).$$

Suppose that there is a bounded function $u(t,x)$ which is C^1 in t and C^2 in x, and a bounded function $f(x)$ in C^2, such that the following PDE is satisfied:

$$\frac{\partial}{\partial t}\,u(t,x) = \mathscr{L}u(t,x), \qquad u(0,x) = f(x).$$

Then $\mathbb{E}(f(X_t)|\mathcal{F}_s) = u(t-s,X_s)$ a.s. for all $0 \le s \le t \le T$, i.e., $u(t,x) = P_t f(x)$.

Remark 5.2.5. The operator $\mathscr{L}$ should look extremely familiar—this is precisely the expression that shows up in Itô's rule! Not surprisingly, this is the key to the proof. $\mathscr{L}$ will show up frequently in the rest of the course.

Remark 5.2.6. You might wonder why the above PDE is called the *backward* equation. In fact, we can just as easily write the equation backwards in time: setting $v(t,x) = u(T-t,x)$ and using the chain rule, we obtain

$$\frac{\partial}{\partial t}\,v(t,x) + \mathscr{L}v(t,x) = 0, \qquad v(T,x) = f(x),$$

which has a terminal condition (at $t = T$) rather than an initial condition (at $t = 0$). For time-nonhomogeneous Markov processes, the latter (backward) form is the appropriate one, so it is in some sense more fundamental. As we have assumed time-homogeneity, however, the two forms are completely equivalent in our case.

Remark 5.2.7. The choice to present proposition 5.2.4 in this way raises a dilemma. In principle the result is "the wrong way around": we would like to use the expression $\mathbb{E}(f(X_t)|\mathcal{F}_s) = u(t-s,X_s)$ to define $u(t,x)$, and then prove that $u(t,x)$ must consequently satisfy the PDE. This is indeed possible in many cases, see [Fri75, theorem 5.6.1]; it is a more technical exercise, however, as we would have to prove that $u(t,x)$ is sufficiently smooth rather than postulating it. More generally, one could prove that this PDE almost always makes sense, in a suitable weak sense, even when $u(t,x)$ is *not* sufficiently differentiable. Though this is theoretically interesting, it is not obvious how to use such a result in applications (are there numerical methods for solving such equations?). We will face this dilemma again when we study optimal control.

After all the remarks, the proof is a bit of an anti-climax.

Proof. Set $v(r,x) = u(t-r,x)$, and apply Itô's rule to $Y_r = v(r,X_r)$. Then we obtain

$$v(t,X_t) = v(0,X_0) + \int_0^t \big[v'(r,X_r) + \mathscr{L}v(r,X_r)\big]\,dr + \text{local martingale}.$$

The time integral vanishes by the Kolmogorov backward equation, so $v(t, X_t)$ is a local martingale. Introducing a localizing sequence $\tau_n \nearrow \infty$, we find using the martingale property

$$\mathbb{E}(v(t \wedge \tau_n, X_{t\wedge\tau_n})|\mathcal{F}_s) = v(s \wedge \tau_n, X_{s\wedge\tau_n}).$$

But as we have assumed that v is bounded, we obtain using dominated convergence for conditional expectations that $\mathbb{E}(f(X_t)|\mathcal{F}_s) = \mathbb{E}(v(t, X_t)|\mathcal{F}_s) = v(s, X_s) = u(t - s, X_s)$. □

Let us now investigate the Kolmogorov *forward* equation, which is in essence the dual of the backward equation. The idea is as follows. For a fixed time t, the random variable X_t is just an $\mathbb{R}^n$-valued random variable. If we are in luck, then the law of this random variable is absolutely continuous with respect to the Lebesgue measure, and, in particular, we can write undergraduate-style

$$\mathbb{E}(f(X_t)) = \int_{\mathbb{R}^n} f(y)\, p_t(y)\, dy$$

with some probability density $p_t(y)$. More generally, we could try to find a *transition density* $p_t(x, y)$ that satisfies for all sufficiently nice f

$$\mathbb{E}(f(X_t)|\mathcal{F}_s) = \int_{\mathbb{R}^n} f(y)\, p_{t-s}(X_s, y)\, dy.$$

The existence of such densities is a nontrivial matter; in fact, there are many reasonable models for which they do not exist. On the other hand, if they were to exist, one can ask whether $p_t(y)$ or $p_t(x, y)$ can again be obtained as the solution of a PDE.

Let us consider, in particular, the (unconditional) density $p_t(y)$. Note that the tower property of the conditional expectation implies that

$$\int_{\mathbb{R}^n} f(y)\, p_t(y)\, dy = \mathbb{E}(f(X_t)) = \mathbb{E}(\mathbb{E}(f(X_t)|X_0)) = \int_{\mathbb{R}^n} P_t f(y)\, p_0(y)\, dy,$$

where $p_0(y)$ is the probability density of X_0 (provided it exists). This explains in what sense the Kolmogorov forward equation is the *dual* of the Kolmogorov backward equation. To prove the forward equation, however, we revert to using Itô's rule.

Proposition 5.2.8 (Kolmogorov forward equation). *Assume that, in addition to the conditions of theorem 5.1.3, $b(x)$ is C^1 and $\sigma(x)$ is C^2. For $\rho \in C^2$, define*

$$\mathscr{L}^*\rho(x) = -\sum_{i=1}^n \frac{\partial}{\partial x^i}(b^i(x)\rho(x)) + \frac{1}{2}\sum_{i,j=1}^n \sum_{k=1}^m \frac{\partial^2}{\partial x^i\, \partial x^j}(\sigma^{ik}(x)\sigma^{jk}(x)\rho(x)).$$

Suppose that the density $p_t(x)$ exists and is C^1 in t, C^2 in x. Then

$$\frac{\partial}{\partial t}\, p_t(x) = \mathscr{L}^* p_t(x), \qquad t \in [0, T],$$

i.e., the density $p_t(x)$ of X_t must satisfy the Kolmogorov forward equation.

Remark 5.2.9. The Kolmogorov forward equation is sometimes referred to as the *Fokker-Planck* equation, particularly in the natural sciences.

Proof. Fix an $f \in C_0^2$ (in C^2 and with compact support). By Itô's rule, we obtain

$$f(X_t) = f(X_0) + \int_0^t \mathscr{L} f(X_s)\, ds + \text{martingale}$$

(the last term is a martingale as f, and hence its derivatives, have compact support, and thus the integrand is bounded). Taking the expectation and using Fubini's theorem, we obtain

$$\mathbb{E}(f(X_t)) = \mathbb{E}(f(X_0)) + \int_0^t \mathbb{E}(\mathscr{L} f(X_s))\, ds.$$

Substituting the definition of $p_t(y)$, integrating by parts, and using Fubini's theorem again,

$$\int_{\mathbb{R}^n} f(y)\, p_t(y)\, dy = \int_{\mathbb{R}^n} f(y)\, p_0(y)\, dy + \int_{\mathbb{R}^n} f(y) \int_0^t \mathscr{L}^* p_s(y)\, ds\, dy.$$

Now note that this expression holds for any $f \in C_0^2$, so we can conclude that

$$\alpha(y) = p_t(y) - p_0(y) - \int_0^t \mathscr{L}^* p_s(y)\, ds = 0$$

for all y, except possibly on some subset with measure zero with respect to the Lebesgue measure. To see this, let $\kappa \in C_0^\infty$ be a nonnegative function such that $\kappa(y) = 1$ for for $\|y\| \le K$. As $\int \alpha(y) f(y) dy = 0$ for any $f \in C_0^2$, we find in particular that

$$\int_{\|y\| \le K} |\alpha(y)|^2\, dy \le \int_{\mathbb{R}^n} \kappa(y) |\alpha(y)|^2\, dy = 0$$

by setting $f(y) = \kappa(y)\alpha(y)$. But then evidently the set $\{y : \|y\| \le K,\ \alpha(y) \neq 0\}$ has measure zero, and as this is the case for any K the claim is established. But $\alpha(y)$ must then be zero everywhere, as it is a continuous in y (this follows by dominated convergence, as $\mathscr{L}^* p_t(y)$ is continuous in (t, y), and hence locally bounded). It remains to take the time derivative. □

Remark 5.2.10. As stated, these theorems are not too useful; the backward equation requires us to show the existence of a sufficiently smooth solution to the backward PDE, while for the forward equation we somehow need to establish that the density of X_t exists and is sufficiently smooth. As a rule of thumb, the backward equation is very well behaved, and will often have a solution provided only that f is sufficiently smooth; the forward equation is much less well behaved and requires stronger conditions on the coefficients b and σ. This is why the backward equation is often more useful as a mathematical tool. Of course, this is only a rule of thumb; a good source for actual results is the book by Friedman [Fri75]. A typical condition for the existence of a smooth density is the *uniform ellipticity* requirement $\sum_{i,j,k} v^i \sigma^{ik}(x) \sigma^{jk}(x) v^j \ge \varepsilon \|v\|^2$ for all $x, v \in \mathbb{R}^n$ and some $\varepsilon > 0$.

5.3 The Wong-Zakai theorem

Even though we have defined stochastic differential equations, and proved the existence and uniqueness of solution, it is not entirely obvious that these mathematical objects really behave like ordinary differential equations. In particular, we would like to think of stochastic differential equations as ordinary differential equations driven

by white noise; but does this actually make sense? The resolution of this problem is important if we want to use SDE to model noise-driven physical systems.

To study this problem, let us start with ordinary differential equations that are driven by rapidly fluctuating—but not white—noise. In particular, define X_t^n to be the solution of the ordinary differential equation

$$\frac{d}{dt}X_t^n = b(X_t^n) + \sigma(X_t^n)\,\xi_t^n, \qquad X_0^n = X_0,$$

where b and σ are Lipschitz continuous as usual, X_0 is a random variable independent of ξ_t^n, and ξ_t^n is some "nice" m-dimensional random process which "approximates" white noise. What does this mean? By "nice", we mean that it is sufficiently smooth that the above equation has a unique solution in the usual ODE sense; to be precise, we will assume that every sample path of ξ_t^n is piecewise continuous. By "approximates white noise", we mean that there is an m-dimensional Wiener process W_t such that

$$\sup_{t\in[0,T]} \|W_t - W_t^n\| \xrightarrow{n\to\infty} 0 \quad \text{a.s.}, \qquad W_t^n = \int_0^t \xi_s^n\,ds.$$

In other words, the time integral of ξ_t^n approximates (uniformly) the Wiener process W_t, which conforms to our intuition of white noise as the derivative of a Wiener process. You can now think of the processes X_t^n as being physically realistic models; on the other hand, these models are almost certainly not Markov, for example. The question is whether when n is very large, X_t^n is well approximated by the solution of a suitable SDE. That SDE is then the corresponding idealized model, which, formally, corresponds to replacing ξ_t^n by white noise.

Can we implement these ideas? Let us first consider the simplest case.

Proposition 5.3.1. *Suppose that $\sigma(x) = \sigma$ does not depend on x, and consider the SDE $dX_t = b(X_t)\,dt + \sigma\,dW_t$. Then $\sup_{t\in[0,T]} \|X_t^n - X_t\| \to 0$ a.s.*

Proof. Note that in this case, we can write

$$X_t^n - X_t = \int_0^t (b(X_s^n) - b(X_s))\,ds + \sigma\,(W_t^n - W_t).$$

Hence we obtain using the triangle inequality and the Lipschitz property

$$\|X_t^n - X_t\| \le K\int_0^t \|X_s^n - X_s\|\,ds + \|\sigma\| \sup_{t\in[0,T]} \|W_t^n - W_t\|.$$

Thus, by Grownwall's lemma, we can write

$$\sup_{t\in[0,T]} \|X_t^n - X_t\| \le e^{KT}\|\sigma\| \sup_{s\in[0,T]} \|W_s^n - W_s\| \xrightarrow{n\to\infty} 0 \quad \text{a.s.}$$

Thus the proof is complete. □

Apparently the processes X_t^n limit to the solution of an SDE, which is precisely the equation that we naively expect, when σ is constant. When $\sigma(x)$ does depend on x, however, we are in for a surprise. For sake of demonstration we will develop this

case only in the very simplest setting, following in the footsteps of Wong and Zakai [WZ65]. This is sufficient to see what is going on and avoids excesssive pain and suffering; a more general result is quoted at the end of the section.

Let us make the following assumptions (even simpler than those in [WZ65]).

1. X_t^n and ξ_t^n are scalar processes (we work in one dimension);
2. b and σ are Lipschitz continuous and bounded;
3. σ is C^1 and $\sigma(x)\, d\sigma(x)/dx$ is Lipschitz continuous; and
4. $\sigma(x) \geq \beta$ for all x and some $\beta > 0$.

The claim is that the solutions of the ODEs

$$\frac{d}{dt} X_t^n = b(X_t^n) + \sigma(X_t^n)\, \xi_t^n, \qquad X_0^n = X_0,$$

converge, as $n \to \infty$, to the solution of the following SDE:

$$dX_t = \left[b(X_t) + \frac{1}{2}\, \sigma(X_t)\, \frac{d\sigma}{dx}(X_t) \right] dt + \sigma(X_t)\, dW_t.$$

By our assumptions, the latter equation still has Lipschitz coefficients and thus has a unique solution. The question is, of course, why the additional term in the time integral (the *Itô correction term*) has suddenly appeared out of nowhere.

Remark 5.3.2. Let us give a heuristic argument for why we expect the Itô correction to be there. Let $f : \mathbb{R} \to \mathbb{R}$ be a *diffeomorphism* (a smooth bijection with smooth inverse). Then setting $Y_t^n = f(X_t^n)$ and using the chain rule gives another ODE

$$\frac{d}{dt} Y_t^n = \frac{df}{dx}(f^{-1}(Y_t^n))\, b(f^{-1}(Y_t^n)) + \frac{df}{dx}(f^{-1}(Y_t^n))\, \sigma(f^{-1}(Y_t^n))\, \xi_t^n.$$

The only thing that has happened here is a (smooth) change of variables. If our limit as $n \to \infty$ is consistent, then it should commute with such a change of variables, i.e., it should not matter whether we first perform a change of variables and then take the white noise limit, or whether we first take the white noise limit and then make a change of variables (after all, we have not changed anything about our system, we have only reparametrized it!). Let us verify that this is indeed the case.

We presume that the limit as $n \to \infty$ works the way we claim it does. Then to obtain the limiting SDE for Y_t^n, we need to calculate the corresponding Itô correction term. This is a slightly gory computation, but some calculus gives

$$\text{Itô corr.} = \frac{1}{2}\sigma(f^{-1}(x))\frac{df}{dx}(f^{-1}(x))\frac{d\sigma}{dx}(f^{-1}(x)) + \frac{1}{2}(\sigma(f^{-1}(x)))^2\frac{d^2 f}{dx^2}(f^{-1}(x)).$$

In particular, we expect that Y_t^n limits to the solution Y_t of the SDE

$$dY_t = \frac{df}{dx}(f^{-1}(Y_t))\, b(f^{-1}(Y_t))\, dt + \frac{df}{dx}(f^{-1}(Y_t))\, \sigma(f^{-1}(Y_t))\, dW_t + \frac{1}{2}\left[\sigma(f^{-1}(Y_t))\frac{df}{dx}(f^{-1}(Y_t))\frac{d\sigma}{dx}(f^{-1}(Y_t)) + (\sigma(f^{-1}(Y_t)))^2\frac{d^2 f}{dx^2}(f^{-1}(Y_t)) \right] dt.$$

But this is precisely the same expression as we would obtain by applying Itô's rule to $Y_t = f(X_t)$! Hence we do indeed find that our limit is invariant under change of variables, precisely as it should be. On the other hand, if we were to neglect to add the Itô correction term, then you can easily verify that this would no longer be the case. In some sense, the Itô correction term "corrects" for the fact that integrals

$$\int_0^t \cdots \, dW_s \qquad \text{and} \qquad \int_0^t \cdots \, \xi_s^n \, ds$$

do not obey the same calculus rules. The additional term in the Itô rule as compared to the ordinary chain rule is magically cancelled by the Itô correction term, thus preventing us from ending up with an unsettling paradox.

That the Itô correction term should cancel the additional term in the Itô rule does not only guide our intuition; this idea is implicitly present in the proof. Notice what happens below when we calculate $\Phi(X_t)$ and $\Phi(X_t^n)$!

Theorem 5.3.3 (Wong-Zakai). $\sup_{t\in[0,T]} |X_t^n - X_t| \to 0$ *a.s. (assuming 1–4 above).*

Proof. Consider the function $\Phi(x) = \int_0^x (\sigma(y))^{-1} dy$, which is well defined and C^2 by the assumption that σ is C^1 and $\sigma(y) \geq \beta > 0$. Then we obtain

$$\frac{d}{dt}\,\Phi(X_t^n) = \frac{b(X_t^n)}{\sigma(X_t^n)} + \xi_t^n, \qquad d\Phi(X_t) = \frac{b(X_t)}{\sigma(X_t)}\,dt + dW_t,$$

using the chain rule and the Itô rule, respectively. In particular, can estimate

$$|\Phi(X_t^n) - \Phi(X_t)| \leq \int_0^t \left| \frac{b(X_s^n)}{\sigma(X_s^n)} - \frac{b(X_s)}{\sigma(X_s)} \right| ds + \sup_{t\in[0,T]} |W_t^n - W_t|.$$

But note that we can write, using the boundedness of σ,

$$|\Phi(x) - \Phi(z)| = \left| \int_x^z \frac{1}{\sigma(y)}\,dy \right| \geq \frac{1}{C_1}\,|x - z|,$$

while using that b is Lipschitz and bounded and that σ is Lipschitz and bounded from below,

$$\left| \frac{b(x)}{\sigma(x)} - \frac{b(z)}{\sigma(z)} \right| \leq \frac{|b(x) - b(z)|}{|\sigma(x)|} + \frac{|b(z)|}{|\sigma(x)\sigma(z)|}|\sigma(z) - \sigma(x)| \leq C_2\,|x - z|.$$

Hence we obtain the estimate

$$|X_t^n - X_t| \leq C_1 C_2 \int_0^t |X_s^n - X_s|\,ds + C_1 \sup_{t\in[0,T]} |W_t^n - W_t|.$$

Applying Gronwall's lemma and taking the limit as $n \to \infty$ completes the proof. □

The result that we have just proved is too restrictive to be of much practical use; however, the lesson learned is an important one. Similar results can be obtained in higher dimensions, with multiple driving noises, unbounded coefficients, etc., at the expense of a large number of gory calculations. To provide a result that is sufficiently general to be of practical interest, let us quote the following theorem from [IW89, theorem VI.7.2] (modulo some natural, but technical, conditions).

Theorem 5.3.4. *Let ξ_t^n be a sequence of approximations to m-dimensional white noise, which are assumed to satisfy a set of conditions [IW89, definition VI.7.1] of a technical nature. Let $b : \mathbb{R}^n \to \mathbb{R}^n$ and $\sigma : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ be C^1 and C^2, respectively, and assume that all their derivatives are bounded. Finally, assume that the initial condition $X_0 \in \mathbb{R}^n$ is non-random. Denote by X_t^n the solutions of*

$$\frac{d}{dt}X_t^n = b(X_t^n) + \sigma(X_t^n)\,\xi_t^n, \qquad X_0^n = X_0,$$

and by X_t the solution of $dX_t = \tilde{b}(X_t)\,dt + \sigma(X_t)\,dW_t$ with the Itô-corrected drift

$$\tilde{b}^i(x) = b^i(x) + \frac{1}{2}\sum_{j=1}^{n}\sum_{k=1}^{m}\frac{\partial\sigma^{ik}(x)}{\partial x^j}\sigma^{jk}(x).$$

Then $\mathbb{E}\left[\sup_{t\in[0,T]}|X_t^n - X_t|^2\right] \xrightarrow{n\to\infty} 0$ for any $T < \infty$.

You have all the tools you need to prove this theorem, provided you have a reliable supply of scratch paper and a pen which is not about to run out of ink. A brief glance at the proof in [IW89] will convince you why it is omitted here.

Remark 5.3.5 (Fisk-Stratonovich integrals). As mentioned previously, the reason for the Itô correction term is essentially that the Itô integral does not obey the ordinary chain rule. This is by no means a conceptual problem; you should simply see the definition of the Itô integral as a mathematical construction, while the theorems in this section justify the modelling of physical phenomena within this framework (and tell us how this should be done properly). However, we have an alternative choice at our disposal: we can choose a different definition for the stochastic integral which *does* obey the chain rule, as was done by Fisk and Stratonovich. When expressed in terms of the Fisk-Stratonovich (FS-)integral, it is precisely the Itô correction which vanishes and we are left with an SDE which looks identical to the ODEs we started with.

There are many problems with the FS-integral, however. First of all, the integral is not a martingale, and its expectation consequently rarely vanishes. This means that this integral is extremely inconvenient in computations that involve expectations. Second, the FS-integral is much less general than the Itô integral, in the sense that the class of stochastic processes which are integrable is significantly smaller than the Itô integrable processes. In fact, the most mathematically sound way to define the FS-integral is as the sum of an Itô integral and a correction term (involving quadratic variations of the integrand and the Wiener process), see [Pro04]. Hence very little is won by using the FS-integral, except a whole bunch of completely avoidable inconvenience. What you win is that the ordinary chain rule holds for the FS-integral, but the Itô rule is just as easy to remember as the chain rule once you know what it is!

For these reasons, we will avoid discussing FS-integrals any further in this course. That being said, however, there is one important case where FS-integrals make more sense than Itô integrals. If we are working on a *manifold* rather than in $\mathbb{R}^n$, the FS-integral can be given an intrinsic (coordinate-free) meaning, whereas this is not true for the Itô integral. This makes the FS-integral the tool of choice for studying stochastic calculus in manifolds (see, e.g., [Bis81]). Itô integrals can also be defined in this setting, but one needs some additional structure: a Riemannian connection.

5.4 The Euler-Maruyama method

Stochastic differential equations, like their non-random counterparts, rarely admit analytical solution. For this reason, it is important to have numerical methods to simulate such equations on a computer. In the SDE case, we are seeking a numerical method that can simulate (approximate) sample paths of the SDE with (approximately) the correct distribution. We will discuss here the simplest of these methods, which is nonetheless one of the most widely used in practice—the *Euler-Maruyama method*.

The method is in fact very close to the classical Euler method for discretization of ODEs. Consider our usual SDE, and let us discretize the interval $[0, T]$ into time steps of length T/p; i.e., we introduce the discrete grid $t_k = kT/p$, $k = 0, \ldots, p$. Then

$$X_{t_n} = X_{t_{n-1}} + \int_{t_{n-1}}^{t_n} b(X_s)\, ds + \int_{t_{n-1}}^{t_n} \sigma(X_s)\, dW_s.$$

This expression can not be used as a numerical method, as X_{t_n} depends not only on $X_{t_{n-1}}$ but on all X_s in the interval $s \in [t_{n-1}, t_n]$. As X_s has continuous sample paths, however, it seems plausible that $X_s \approx X_{t_{n-1}}$ for $s \in [t_{n-1}, t_n]$, provided that p is sufficiently large. Then we can try to approximate

$$X_{t_n} \approx X_{t_{n-1}} + \int_{t_{n-1}}^{t_n} b(X_{t_{n-1}})\, ds + \int_{t_{n-1}}^{t_n} \sigma(X_{t_{n-1}})\, dW_s,$$

or, equivalently,

$$X_{t_n} \approx X_{t_{n-1}} + b(X_{t_{n-1}})\,(t_n - t_{n-1}) + \sigma(X_{t_{n-1}})\,(W_{t_n} - W_{t_{n-1}}).$$

This simple recursion is easily implemented on a computer, where we can obtain a suitable sequence $W_{t_n} - W_{t_{n-1}}$ by generating i.i.d. m-dimensional Gaussian random variables with mean zero and covariance $(T/p)I$ using a (pseudo-)random number generator. The question that we wish to answer is whether this algorithm really does approximate the solution of the full SDE when p is sufficiently large.

The remainder of this section is devoted to proving convergence of the Euler-Mayurama method. Before we proceed to the proof of that result, we need a simple estimate on the increments of the solution of an SDE.

Lemma 5.4.1. *Under the assumptions of theorem 5.1.3, we can estimate*

$$\|X_t - X_s\|_{2,\mathbb{P}} \le L\sqrt{t-s}, \qquad 0 \le s \le t \le T,$$

where the constant L depends only on T, X_0, b and σ.

Proof. The arguments are similar to those used in the proof of theorem 5.1.3. Write

$$\mathbb{E}(\|X_t - X_s\|^2) \le 2\,\mathbb{E}\left(\left\|\int_s^t b(X_r)\, dr\right\|^2\right) + 2\,\mathbb{E}\left(\left\|\int_s^t \sigma(X_r)\, dW_r\right\|^2\right),$$

where we have used the identity $(a+b)^2 \le 2(a^2+b^2)$. But

$$\mathbb{E}\left(\left\|\int_s^t b(X_r)\, dr\right\|^2\right) \le (t-s)\int_s^t \mathbb{E}(\|b(X_r)\|^2)\, dr \le 2CT\int_s^t \mathbb{E}(1+\|X_r\|^2)\, dr,$$

using the linear growth condition and the same identity for $(a+b)^2$. Similarly,

$$\mathbb{E}\left(\left\|\int_s^t \sigma(X_r)\,dW_r\right\|^2\right) = \int_s^t \mathbb{E}(\|\sigma(X_r)\|^2)\,dr \leq 2C\int_s^t \mathbb{E}(1+\|X_r\|^2)\,dr,$$

using the Itô isometry. But it was established in the proof of theorem 5.1.5 that $\mathbb{E}(1+\|X_r\|^2)$ is bounded by a constant that only depends on X_0, T and C. Hence the result follows. □

We will also need a discrete version of Gronwall's lemma.

Lemma 5.4.2 (Discrete Gronwall). *Let $A, B > 0$ and let $\alpha_n \geq 0$, $n = 0, \ldots, N$. If*

$$\alpha_n \leq A + B\sum_{k=1}^{n} \alpha_{k-1}, \qquad n = 0, \ldots, N,$$

then it must be the case that $\alpha_n \leq Ae^{Bn}$ for all $0 \leq n \leq N$.

Proof. Suppose that we have established that $\alpha_k \leq Ae^{Bk}$ for all $0 \leq k \leq n-1$. Then

$$\alpha_n \leq A + AB\sum_{k=1}^{n} e^{B(k-1)} = A + AB\,\frac{e^{Bn}-1}{e^B-1} \leq Ae^{Bn},$$

where we have used $e^B \geq 1+B$. But $\alpha_0 \leq A = Ae^{B\cdot 0}$, so the result follows by induction. □

Let us now complete the proof of convergence of the Euler-Maruyama scheme as $p \to \infty$. The approximate solution is defined recursively as

$$Y_{t_n} = Y_{t_{n-1}} + b(Y_{t_{n-1}})\,(t_n - t_{n-1}) + \sigma(Y_{t_{n-1}})\,(W_{t_n} - W_{t_{n-1}}),$$

and we wish to prove that Y_{t_n} is close to X_{t_n} in a suitable sense.

Theorem 5.4.3 (Order 0.5 convergence of the Euler-Maruyama method). *Assume, in addition to the assumptions of theorem 5.1.3, that Y_0 is chosen in such a way that $\|X_0 - Y_0\|_{2,\mathbb{P}} \leq C_1 p^{-1/2}$ for some constant $C_1 < \infty$. Then*

$$\max_{0\leq n\leq p} \|X_{t_n} - Y_{t_n}\|_{2,\mathbb{P}} \leq C_2 p^{-1/2},$$

where $C_2 < \infty$ is a constant that depends only on T, X_0, C_1, b and σ.

Proof. Define the process $Y_t = Y_{t_{k-1}}$ for $t \in [t_{k-1}, t_k[$. Then

$$X_{t_n} - Y_{t_n} = X_0 - Y_0 + \int_0^{t_n} (b(X_s) - b(Y_s))\,ds + \int_0^{t_n} (\sigma(X_s) - \sigma(Y_s))\,dW_s.$$

Hence we obtain, using the triangle inequality and $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$,

$$\|X_{t_n} - Y_{t_n}\|^2 \leq 3\,\|X_0 - Y_0\|^2 + 3\left\|\int_0^{t_n} (b(X_s) - b(Y_s))\,ds\right\|^2 + 3\left\|\int_0^{t_n} (\sigma(X_s) - \sigma(Y_s))\,dW_s\right\|^2.$$

Taking the expectation and proceeding as in the proof of theorem 5.1.3, we can estimate

$$\mathbb{E}(\|X_{t_n} - Y_{t_n}\|^2) \leq 3\,\mathbb{E}(\|X_0 - Y_0\|^2) + 3\,t_n \int_0^{t_n} \mathbb{E}(\|b(X_s) - b(Y_s)\|^2)\,ds + 3\int_0^{t_n} \mathbb{E}(\|\sigma(X_s) - \sigma(Y_s)\|^2)\,ds.$$

Now use the Lipschitz continuity of b and σ: this gives

$$\mathbb{E}(\|X_{t_n} - Y_{t_n}\|^2) \leq 3\,\mathbb{E}(\|X_0 - Y_0\|^2) + 3K^2(T+1)\int_0^{t_n} \mathbb{E}(\|X_s - Y_s\|^2)\,ds.$$

At this point we need to estimate $\mathbb{E}(\|X_s - Y_s\|^2)$. Suppose that $s \in [t_{k-1}, t_k[$; then

$$\|X_s - Y_s\|_{2,\mathbb{P}} = \|X_s - Y_{t_{k-1}}\|_{2,\mathbb{P}} \leq \|X_s - X_{t_{k-1}}\|_{2,\mathbb{P}} + \|X_{t_{k-1}} - Y_{t_{k-1}}\|_{2,\mathbb{P}}.$$

Using the lemma 5.4.1 and $s - t_{k-1} \leq T/p$, we obtain

$$\mathbb{E}(\|X_s - Y_s\|^2) \leq \frac{2L^2T}{p} + 2\,\mathbb{E}(\|X_{t_{k-1}} - Y_{t_{k-1}}\|^2).$$

Thus we can now write, using the definition of C_1,

$$\mathbb{E}(\|X_{t_n} - Y_{t_n}\|^2) \leq \frac{3C_1^2 + 6K^2L^2T^2(T+1)}{p} + \frac{6K^2T(T+1)}{p}\sum_{k=1}^{n} \mathbb{E}(\|X_{t_{k-1}} - Y_{t_{k-1}}\|^2).$$

If we set $J_n = \max_{0\leq i\leq n} \mathbb{E}(\|X_{t_i} - Y_{t_i}\|^2)$, then we can evidently write

$$J_n \leq \frac{C_3}{p} + \frac{C_4T}{p}\sum_{k=1}^{n} J_{k-1},$$

where we have replaced some of the unsightly expressions by friendly-looking symbols. But we can now apply the discrete Gronwall lemma 5.4.2 to obtain $J_n \leq (C_3/p)\exp(C_4Tn/p)$. Hence the result follows if we define $C_2 = \sqrt{C_3}\,\exp(C_4T/2)$. □

5.5 Stochastic stability

In non-random nonlinear systems and control theory, the notions of Lyapunov stability and Lyapunov functions play an important role (see, e.g., [Kha02]). Let us briefly recall the most basic concepts in this theory. The starting point is an ODE of the form

$$\frac{dX(t)}{dt} = b(X(t)), \qquad X(0) \in \mathbb{R}^n,$$

where $b(x)$ vanishes at some point $x^* \in \mathbb{R}^n$. The point x^* is called an *equilibrium* of the ODE, because if we start at $X(0) = x^*$, then $X(t) = x^*$ for all t (note that there may be multiple equilibria). The question is, if we start close to x^* rather than on x^*, whether we will always remain close, or, better even, whether we will converge to x^* as $t \to \infty$. It is this type of question that is addressed by the Lyapunov theory.

The formal definitions are as follows. The equilibrium position x^* is said to be

- **stable** if for any $\varepsilon > 0$, there is a $\delta > 0$ such that $\|X(t) - x^*\| < \varepsilon$ for all $t \geq 0$ whenever $\|X(0) - x^*\| < \delta$;

- **asymptotically stable** if it is stable and there is a $\kappa > 0$ such that $X(t) \to x^*$ as $t \to \infty$ whenever $\|X(0) - x^*\| < \kappa$; and
- **globally stable** if it is stable and $X(t) \to x^*$ as $t \to \infty$ for any $X(0) \in \mathbb{R}^n$.

In other words, the equilibrium x^* is stable if we are guaranteed to remain close to x^* forever provided that we start sufficiently close, is asymptotically stable if we are additionally guaranteed to converge to x^* if we start sufficiently close, and is globally stable if we always converge to x^* no matter where we started.

Can we study such problems in the stochastic case? There are various interesting questions that we can ask, but they depend on the form of the SDE. For example, a common way to add stochastic perturbations to an ODE is through additive noise:

$$dX_t = b(X_t)\,dt + \varepsilon\,dW_t, \qquad X_0 = X(0).$$

Even if $b(x^*) = 0$, the process X_t will not remain at x^* even if we start there: the noise will kick us away from the deterministic equilibrium. However, one of the justifications for studying deterministic stability is that an asymptotically stable equilibrium point should be robust against perturbations. Thus we expect that if we add a small perturbing noise—i.e., $\varepsilon \ll 1$—then, even though we will not remain at x^*, we will be very likely to find ourselves close to x^* at any time in the future.[2] There is a simple type of result that can help quantify this idea. [Note that we have only chosen $\sigma(x) = \varepsilon$ for sake of demonstration; the following result holds for arbitrary $\sigma(x)$.]

Proposition 5.5.1. *Assume that the conditions of theorem 5.1.3 hold, and define $\mathscr{L}$ as in proposition 5.2.4. Suppose that there exists a function $V : \mathbb{R}^n \to [0, \infty[$ which is C^2 and satisfies $\mathscr{L}V(x) \le -\alpha V(x) + \beta$ for all $x \in \mathbb{R}^n$ and some $\alpha, \beta > 0$. Then*

$$\mathbb{E}(V(X_t)) \le e^{-\alpha t}\,\mathbb{E}(V(X_0)) + \frac{\beta}{\alpha}, \qquad \forall\, t \ge 0,$$

provided that $\mathbb{E}(V(X_0)) < \infty$.

Remark 5.5.2. Suppose, for example, that we choose the function $V(x)$ such that $V(x) \ge \|x - x^*\|^p$ for some $p > 0$. Then $V(x)$ is a measure of the distance to the equilibrium point, and this result bounds the expected distance from the equilibrium point uniformly in time. In particular, using Chebyshev's inequality, you can obtain a bound on the probability of being far from equilibrium at any fixed time.

Proof. Using Itô's rule, we obtain immediately

$$V(X_t)\,e^{\alpha t} = V(X_0) + \int_0^t e^{\alpha s}(\mathscr{L}V(X_s) + \alpha V(X_s))\,ds + \text{local martingale}.$$

Let $\tau_n \nearrow \infty$ be a localizing sequence. Then we have

$$\mathbb{E}(V(X_{t\wedge\tau_n})\,e^{\alpha\,t\wedge\tau_n}) = \mathbb{E}(V(X_0)) + \mathbb{E}\left[\int_0^{t\wedge\tau_n} e^{\alpha s}(\mathscr{L}V(X_s) + \alpha V(X_s))\,ds\right].$$

[2] For the particular case of small noise, there is a powerful theory to study the asymptotic properties of SDEs as $\varepsilon \to 0$: the *Freidlin-Wentzell theory of large deviations* [FW98]. Unfortunately, we will not have the time to explore this interesting subject. The theorems in this section are fundamentally different; they are not asymptotic in nature, and work for any SDE (provided a suitable function V can be found!).

Now use $\mathscr{L}V(x) + \alpha V(x) \leq \beta$ to obtain

$$\mathbb{E}(V(X_t)\, e^{\alpha t}) \leq \mathbb{E}(V(X_0)) + \beta \int_0^t e^{\alpha s}\, ds,$$

where we have used Fatou's lemma and monotone convergence to take the limit as $n \to \infty$ on the left- and right-hand sides, respectively. The conclusion of the result is straightforward. □

A very different situation is one where for the SDE with non-random X_0

$$dX_t = b(X_t)\, dt + \sigma(X_t)\, dW_t, \qquad X_0 = x,$$

both $b(x)$ and $\sigma(x)$ vanish simultaneously at $x = x^*$. Then $X_t = x^*$ for all $t \geq 0$ is certainly a solution of the SDE, so x^* is a true equilibrium point of the system. One can now develop a true counterpart of the Lyapunov theory for this case. However, expecting that if we start close enough to x^*, we will (for example) converge to x^* with probability one, is often too much to ask: even though the noise vanishes at x^*, the noise is non-negligible outside x^* and might well kick us far away from x^* with some very small probability. It is more fruitful to ask whether we can make X_t converge to x^* with probability close to one, if we start with X_0 sufficiently close to x^*. These considerations motivate the following definitions. The equilibrium x^* is

- **stable** if for any $\varepsilon > 0$ and $\alpha \in]0, 1[$, there exists a $\delta > 0$ such that we have $\mathbb{P}(\sup_{t \geq 0} \|X_t - x^*\| < \varepsilon) > \alpha$ whenever $\|X_0 - x^*\| < \delta$;
- **asymptotically stable** if it is stable and for every $\alpha \in]0, 1[$, there exists a $\kappa > 0$ such that $\mathbb{P}(X_t \to x^* \text{ as } t \to \infty) > \alpha$ whenever $\|X_0 - x^*\| < \kappa$; and
- **globally stable** if it is stable and $X_t \to x^*$ a.s. as $t \to \infty$ for any X_0.

Remark 5.5.3. It is important to understand the distinction between a statement such as $\mathbb{P}(\sup_{t \geq 0} \|X_t - x^*\| \geq \varepsilon) \leq 1 - \alpha$, compared to the *much weaker* statement $\sup_{t \geq 0} \mathbb{P}(\|X_t - x^*\| \geq \varepsilon) \leq 1 - \alpha$ which can be obtained from proposition 5.5.1 using Chebyshev's inequality. The former expresses the fact that the probability that our sample paths will *ever* venture farther away from x^* than a distance ε is very small. The latter, however, expresses the fact that at any fixed time t, the fraction of the sample paths that are farther away from x^* than a distance ε is small. In the latter case, it is quite likely that *every* sample path ventures very far away from x^* at some point in time; they just do not all do so at the same time.

Let us find some simple conditions for the stability of x^*. We always work under the assumptions of theorem 5.1.3 and with the *non-random* initial condition X_0.

Proposition 5.5.4. *Suppose that there exists a function $V : \mathbb{R}^n \to [0, \infty[$ which is C^2 and satisfies $V(x^*) = 0$, $V(x) > 0$ if $x \neq x^*$, and $\mathscr{L}V(x) \leq 0$, for all x in some neighborhood U of x^*. Then x^* is a stable equilibrium for X_t.*

Proof. We wish to prove that we can make $\mathbb{P}(\sup_{t \geq 0} \|X_t - x^*\| \geq \varepsilon)$ arbitrarily small if we choose X_0 sufficiently close to x^*. Note that it suffices to prove this for sufficiently small ε; after all, if the statement holds for any $\varepsilon \leq \varepsilon^*$, then for $\varepsilon > \varepsilon^*$ we can use the trivial inequality

$\mathbb{P}(\sup_{t\geq 0}\|X_t - x^*\| \geq \varepsilon) \leq \mathbb{P}(\sup_{t\geq 0}\|X_t - x^*\| \geq \varepsilon^*)$ to conclude the result. Note that it also suffices to assume that $X_0 \in U$, as this is always the case for X_0 sufficiently close to x^*. Moreover, we can assume that U has compact closure $\overline{U}$, and that $V(x) > 0$ for $x \in \overline{U}\backslash\{x^*\}$; otherwise we can always find an $U' \subset U$ for which this is the case, and proceed with that.

Define $\tau = \inf\{t : X_t \notin U\}$. Using Itô's rule, we obtain

$$V(X_{t\wedge\tau}) = V(X_0) + \int_0^{t\wedge\tau} \mathscr{L}V(X_s)\,ds + \text{martingale}$$

(the stochastic integral stopped at τ is a martingale, as the integrand is bounded for $X_s \in U$). But as $\mathscr{L}V(x) \leq 0$ for $x \in U$, the time integral is nonincreasing with t. Hence $V(X_{t\wedge\tau})$ is a supermartingale, and we get using the supermartingale inequality

$$\mathbb{P}\left[\sup_{t\geq 0} V(X_{t\wedge\tau}) \geq \alpha\right] \leq \frac{V(X_0)}{\alpha}.$$

We now claim that for every $\varepsilon > 0$, there exists an $\alpha > 0$ such that $\|x - x^*\| \geq \varepsilon$ implies $V(x) \geq \alpha$ (for $x \in U$); indeed, just choose α to be the minimum of $V(x)$ over the compact set $\{x \in \overline{U} : \|x - x^*\| \geq \varepsilon\}$ (which is nonempty for sufficiently small ε), and this minimum is strictly positive by our assumptions on V. Hence for any $\varepsilon > 0$, there is an $\alpha > 0$ such that

$$\mathbb{P}\left[\sup_{t\geq 0}\|X_{t\wedge\tau} - x^*\| \geq \varepsilon\right] \leq \frac{V(X_0)}{\alpha},$$

and term on the right can be made arbitrarily small by choosing X_0 sufficiently close to x^*. Finally, it remains to note that $\sup_{t\geq 0}\|X_t - x^*\| \geq \varepsilon$ implies $\sup_{t\geq 0}\|X_{t\wedge\tau} - x^*\| \geq \varepsilon$ if ε is sufficiently small that $\|x - x^*\| \leq \varepsilon$ implies that $x \in U$. □

With almost the same condition, we obtain asymptotic stability.

Proposition 5.5.5. *Suppose that there exists a function $V : \mathbb{R}^n \to [0,\infty[$ which is C^2 and satisfies $V(x^*) = 0$, $V(x) > 0$ if $x \neq x^*$, and $\mathscr{L}V(x) < 0$ if $x \neq x^*$, for all x in some neighborhood U of x^*. Then x^* is asymptotically stable.*

Proof. The current proof is a continuation of the previous proof. Note that

$$\mathbb{E}\left[\int_0^{t\wedge\tau} (-\mathscr{L}V)(X_s)\,ds\right] = V(X_0) - \mathbb{E}(V(X_{t\wedge\tau})) \leq V(X_0) < \infty.$$

But the term on the left is nonnegative and nondecreasing by our assumptions, so we obtain

$$\mathbb{E}\left[\int_0^{\tau} (-\mathscr{L}V)(X_s)\,ds\right] \leq V(X_0) < \infty$$

by monotone convergence. In particular, we find that

$$\int_0^{\tau} (-\mathscr{L}V)(X_s)\,ds < \infty \quad \text{a.s.}$$

If $\tau = \infty$ for some sample path ω, then we conclude that at least $\liminf_{s\to\infty}(-\mathscr{L}V)(X_s) = 0$ for that sample path (except possibly in a set of measure zero). But by our assumption that $\mathscr{L}V(x) < 0$ for $x \neq x^*$, an entirely parallell argument to the one used in the previous proof establishes that $\tau = \infty$ implies $\liminf_{s\to\infty}\|X_s - x^*\| = 0$ and even $\liminf_{s\to\infty} V(X_s) = 0$.

On the other hand, as $V(X_{t\wedge\tau})$ is a nonnegative supermartingale, the martingale convergence theorem holds and we find that $V(X_{t\wedge\tau}) \to Y$ a.s. as $t \to \infty$ for some random variable Y. But then we conclude that for those sample paths (modulo a null set) where $\tau = \infty$, it must be the case that $Y = 0$. In other words, we have established that almost every sample path that stays in U forever must converge to x^*. It remains to note that by the fact that x^* is also stable (which follows by the previous result), we can make the probability that X_t stays in U forever arbitrarily large by starting sufficiently close to x^*. Hence asymptotic stability follows. □

Finally, let us obtain a condition for global stability. The strategy should look a little predictable by now, and indeed there is nothing new here; we only need to assume that our function V is radially unbounded to be able to conclude that $V(X_t) \to 0$ implies $X_t \to x^*$ (as we are no longer working in a bounded neighborhood).

Proposition 5.5.6. *Suppose there exists $V : \mathbb{R}^n \to [0,\infty[$ which is C^2 and satisfies $V(x^*) = 0$, $V(x) > 0$ and $\mathscr{L}V(x) < 0$ for any $x \in \mathbb{R}^n$ such that $x \neq x^*$. Moreover, suppose $V(x) \to \infty$ and $|\mathscr{L}V(x)| \to \infty$ as $\|x\| \to \infty$. Then x^* is globally stable.*

Proof. Using Itô's rule we obtain, by choosing a suitable localizing sequence $\tau_n \nearrow \infty$,

$$\mathbb{E}\left[\int_0^{t\wedge\tau_n} (-\mathscr{L}V)(X_s)\,ds\right] = V(X_0) - \mathbb{E}(V(X_{t\wedge\tau_n})) \leq V(X_0) < \infty,$$

Using monotone convergence, we can send $t \to \infty$ and $n \to \infty$ to conclude that

$$\int_0^\infty (-\mathscr{L}V)(X_s)\,ds < \infty \quad \text{a.s.}$$

But using the fact that $|\mathscr{L}V(x)| \to \infty$ as $\|x\| \to \infty$, we find that $\liminf_{s\to\infty} V(X_s) = 0$ a.s. On the other hand, by Itô's rule, we find that $V(X_t)$ is the sum of a nonincreasing process and a nonnegative local martingale. But then $V(X_t)$ is a nonnegative supermartingle, and the martingale convergence theorem applies. Thus $V(X_t) \to 0$ a.s. It remains to note that we can conclude that $X_t \to x^*$ a.s. using the fact that $V(x) \to \infty$ as $\|x\| \to \infty$. □

Example 5.5.7. Consider the controlled linear stochastic differential equation

$$dX_t = AX_t\,dt + Bu_t\,dt + \sum_{i=1}^k C^i X_t\,dW_t^i, \qquad X_0 \in \mathbb{R}^n,$$

where A is an $n\times n$ matrix, B is an $n\times k$ matrix, C^i are $n\times n$ matrices ($i = 1,\ldots,m$), u_t is a k-dimensional control input and W_t is an m-dimensional Wiener process. We would like to find a linear feedback control strategy $u_t = DX_t$ (D is a $k \times n$ matrix) such that the equilibrium point $x = 0$ is globally stable.

Let us try a Lyapunov function of the form $V_R(x) = x^*Rx$, where R is a positive definite $n \times n$ matrix. Then $V_R(x) = 0$ for $x = 0$, $V_R(x) > 0$ for $x \neq 0$, and $V_R(x) \to \infty$ and $\|x\| \to \infty$. We can now calculate

$$\mathscr{L}V_R(x) = x^*\left[R(A+BD) + (A+BD)^*R + \sum_{i=1}^k (C^i)^*RC^i\right]x \equiv -x^*V[D,R]x.$$

Evidently a sufficient condition for D to be a stabilizing controller is the existence of a positive definite matrix R such that the matrix $V[D,R]$ is positive definite.

Much more can be said about stochastic stability; see section 5.7 for references.

5.6 Is there life beyond the Lipschitz condition?

The Lipschitz condition has played an important role throughout this chapter; but is it truly necessary? The answer is no, but if we are not careful either existence or uniqueness may fail. This is even the case for ordinary differential equations, as the following illuminating examples (from [Øks03]) demonstrate. Consider the ODE

$$\frac{d}{dt}X(t) = 3\,(X(t))^{2/3}, \qquad X(0) = 0.$$

Then $X(t) = (t-a)^3 \vee 0$ is a perfectly respectable solution for any $a > 0$. Evidently, this equation has many solutions for the same initial condition, so uniqueness fails! On the other hand, consider the ODE

$$\frac{d}{dt}X(t) = (X(t))^2, \qquad X(0) = 1.$$

This equation is satisfied only by $X(t) = (1-t)^{-1}$ for $t < 1$, but the solution blows up at $t = 1$. Hence a solution does not exist if we are interested, for example, in the interval $t \in [0, 2]$. Note that neither of these examples satisfy the Lipschitz condition.

There is a crucial difference between these two examples, however. In the first example, the Lipschitz property fails at $x = 0$. On the other hand, in the second example the Lipschitz property fails as $x \to \infty$, but in any compact set the Lipschitz property still holds. Such a function is called *locally* Lipschitz continuous.

Definition 5.6.1. $f : \mathbb{R}^n \to \mathbb{R}^m$ is called *locally Lipschitz continuous* if for any $r < \infty$, there is a $K_r < \infty$ such that $\|f(x)-f(y)\| \le K_r\|x-y\|$ for all $\|x\|, \|y\| \le r$.

For locally Lipschitz coefficents, we have the following result.

Theorem 5.6.2. *Suppose that b and σ are locally Lipschitz continuous. Then the SDE*

$$X_t = X_0 + \int_0^t b(X_s)\,ds + \int_0^t \sigma(X_s)\,dW_s,$$

has a unique solution in the time interval $[0, \zeta[$, where the stopping time ζ is called the explosion time (ζ may be ∞ with positive probability).

Remark 5.6.3. A similar result holds with time-dependent coefficients; we restrict ourselves to the time-homogeneous case for notational simplicity only.

Proof. For any $r < \infty$, we can find functions $b_r(x)$ and $\sigma_r(x)$ which are (globally) Lipschitz and such that $b(x) = b_r(x)$ and $\sigma(x) = \sigma_r(x)$ for all $\|x\| \le r$. For the SDE with coefficients b_r and σ_r and the initial condition $X_0(r) = X_0 I_{\|X_0\| \le r}$, we can find a unique solution $X_t(r)$ for all $t \in [0, \infty[$ using theorem 5.1.3 (and by trivial localization). Now denote by $\tau_r = I_{\|X_0\| \le r} \inf\{t : X_t(r) \ge r\}$, and note that this is a stopping time. Moreover, the process $X_{t\wedge\tau_r}(r)$ evidently satisfies the SDE in the statement of the theorem for $t < \tau_r$. Hence we obtain a unique solution for our SDE in the interval $[0, \tau_r]$. But we can do this for any $r < \infty$, so letting $r \to \infty$ we obtain a unique solution in the interval $[0, \zeta[$ with $\zeta = \lim_{r\to\infty} \tau_r$. □

The proof of this result is rather telling; in going from global Lipschitz coefficients to local Lipschitz coefficients, we proceed as we have done so often by introducing a localizing sequence of stopping time and constructing the solution up to every stopping time. Unlike in the case of the Itô integral, however, these stopping times may accumulate—and we end up with an explosion at the accumulation point.

All is not lost, however: there are many SDEs whose coefficients are only locally Lipschitz, but which nonetheless do not explode! Here is one possible condition.

Proposition 5.6.4. *If $\|X_0\|_{2,\mathbb{P}} < \infty$, b and σ are locally Lipschitz continuous and satisfy a linear growth condition, then the explosion time $\zeta = \infty$ a.s.*

Recally that for Lipschitz coefficients, the linear growth condition follows (see the proof of theorem 5.1.3). In the local Lipschitz setting this is not the case, however, and we must impose it as an additional condition (evidently with desirable results!)

Proof. Proceeding as in the proof of theorem 5.1.5, we find that $\mathbb{E}(\|X_{t\wedge\zeta}\|^2) < \infty$ for all $t < \infty$. But then $X_{t\wedge\zeta} < \infty$ a.s. for all $t < \infty$, so $\zeta = \infty$ a.s. (as $X_\zeta = \infty$ by definition!). □

Remark 5.6.5. All of the conditions which we have discussed for the existence and uniqueness of solutions are only sufficient, but not necessary. Even an SDE with very strange coefficients may have a unique, non-exploding solution; but if it does not fall under any of the standard categories, it might take some specialized work to prove that this is indeed the case. An example of a useful SDE that is not covered by our theorems is the Cox-Ingersoll-Ross equation for the modelling of interest rates:

$$dX_t = (a - bX_t)\,dt + \sigma\sqrt{|X_t|}\,dW_t, \qquad X_0 > 0,$$

with $a, b, \sigma > 0$. Fortunately, however, many (if not most) SDEs which are encountered in applications have at least locally Lipschitz coefficients.

There is an entirely different concept of what it means to obtain a solution of a stochastic differential equation, which we will now discuss very briefly. Let us consider the simplest example: we wish to find a solution of the SDE

$$X_t = \int_0^t b(X_s)\,ds + W_t,$$

where b is some bounded measurable function. Previously, we considered W_t as being a given Wiener process, and we sought to find the solution X_t with respect to this particular Wiener process. This is called a *strong solution*. We can, however, ask a different question: if we do not start from a fixed Wiener process, can we construct (on some probability space) both a Wiener process W_t and a process X_t simultaneously such that the above equation holds? If we can do this, then the solution is called a *weak solution*. Surprisingly, we can always find a weak solution of the above equation—despite the fact that we have imposed almost no structure on b!

Let us perform this miracle. We start with some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, on which is defined a Wiener process X_t. *Note that X_t is now the Wiener process!* Next, we perform a cunning trick. We introduce a new measure $\mathbb{Q}$ as follows:

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = \exp\left(\int_0^t b(X_t)\,dX_t - \frac{1}{2}\int_0^t (b(X_t))^2\,dt\right).$$

This is a Girsanov transformation (and Novikov's condition is satisfied as we have assumed that b is bounded), so we find that the process

$$W_t = X_t - \int_0^t b(X_s)\,ds$$

is a Wiener process under $\mathbb{Q}$. But now we are done: we have constructed a process X_t and a Wiener process W_t on the space $(\Omega, \mathcal{F}, \mathbb{Q})$, so that the desired SDE is satisfied!

You might well be regarding this story with some amount of suspicion—where is the catch? If we fix in hindsight the Wiener process which we have just constructed, and ask for a solution with respect to that Wiener process, then can we not regard X_t as a strong solution with respect to W_t? There is a subtle but very important reason why this is not the case. When we constructed strong solutions, we found that the solution X_t was a functional of the driving noise: a strong solution X_t is $\mathcal{F}_t^W = \sigma\{W_s : s \le t\}$ measurable. This is precisely what you would expect from the point of view of causality: the noise drives a physical system, and thus the state of the physical system is a functional of the realization of the noise. On the other hand, if you look carefully at the construction of our weak solution, you will find precisely the opposite conclusion: that the noise W_t is $\mathcal{F}_t^X = \sigma\{X_s : s \le t\}$ measurable. Evidently, for a weak solution the noise is a functional of the solution of the SDE. Thus it appears that causality is reversed in the weak solution case.

For this reason, you might want to think twice before using weak solutions in modelling applications; the concept of a weak solution is much more probabilistic in nature, while stong solutions are much closer to the classical notion of a differential equation (as our existence and uniqueness proofs, the Wong-Zakai theorem, and the Euler-Maruyama method abundantly demonstrate). Nonetheless weak solutions are an extremely valuable technical tool, both for mathematical purposes and in applications where the existence of solutions in a strong sense may be too restrictive or difficult to verify. Of course, many weak solutions are also strong solutions, so the dilemma only appears if it turns out that a strong solution does not exist.

5.7 Further reading

The recommended texts on stochastic differential equations are, once again, the usual suspects: Øksendal [Øks03] and Arnold [Arn74] for an accessible introduction, and the books by Rogers and Williams [RW00b], Karatzas and Shreve [KS91], Friedman [Fri75], Liptser and Shiryaev [LS01a], or Protter [Pro04] for the Real Thing. Our treatment of existence and uniqueness is inspired by the treatment in Gikhman and Skorokhod [GS96] and to a lesser extent by Ikeda and Watanabe [IW89].

For the general theory of Markov processes, you might want to look in Rogers and Williams [RW00a, chapter III] for a friendly introduction. The classic reference remains Dynkin [Dyn06], and a modern tome is the book by Ethier and Kurtz [EK86]. Friedman [Fri75] is an excellent source on the relation between SDEs and PDEs.

The Wong-Zakai theorem has its origins in Wong and Zakai [WZ65] and was subsequently investigated by various authors (notably the support theorem of Stroock and Varadhan [SV72]). A nice review article is the one by Twardowska [Twa96].

The ultimate bible on numerical methods for stochastic differential equations is the book by Kloeden and Platen [KP92]; there you will find almost any variant of numerical approximation for SDE known to man, at least at the time of publication of that work. Needless to say you can do better than the Euler-Maruyama method (but nonetheless, that simple method is often not too bad!) Our treatment was loosely inspired by lecture notes of Stig Larsson [Lar05]. An intriguing and entirely different way to simulate sample paths of an SDE was recently proposed by Beskos and Roberts [BR05]; they see the solution of an SDE as a path-valued random variable, and use Monte Carlo sampling techniques to sample from its distribution. This is much closer to the weak solution concept than to strong solutions.

Excellent sources for stochastic stability theory are the textbooks by Has'minskii [Has80] and by Kushner [Kus67]. An article by Kushner [Kus72] develops a counterpart of the LaSalle invariance principle in the stochastic setting. The theory of stochastic stability has its origins, in discrete time, in the work of Bucy, see [BJ87], and see also [Kus71] for more discrete time stability theory. Some recent work (also in connection with control) can be found in Deng, Krstić and Williams [DKW01].

Beside the Wentzell-Freidlin large deviations theory [FW98], an omission from this chapter is a study of the dependence of the solution of an SDE on the initial condition. In particular, it is well known that non-random ODEs generate much more than an individual solution for each initial condition: they generate a *flow*, i.e., an entire diffeomorphism of the state space which corresponds to the solution with a particular initial condition at every point. A parallel theory exists for stochastic differential equations, as is detailed, e.g., in the book by Kunita [Kun90]. The most accessible place to start reading are Kunita's lecture notes [Kun84].

CHAPTER 6

Optimal Control

Stochastic optimal control is a highly technical subject, much of which centers around mathematical issues of existence and regularity and is not directly relevant from an engineering perspective. Nonetheless the theory has a large number of applications, many (but not all) of which revolve around the important linear case. In this course we will avoid almost all of the technicalities by focusing on the so-called "verification theorems", which we will encounter shortly, instead of on the more mathematical aspects of the theory. Hopefully this will make the theory both accessible and useful; in any case, it should give you enough ideas to get started.

6.1 Stochastic control problems and dynamic programming

Controlled stochastic differential equations

As usual, we work on $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, \mathbb{P})$ with an m-dimensional $\mathcal{F}_t$-Wiener process W_t. The basic object of interest in stochastic control theory is a stochastic differential equation with a control input: i.e., the state of the controlled system is described by

$$dX_t^u = b(t, X_t^u, u_t)\,dt + \sigma(t, X_t^u, u_t)\,dW_t, \qquad X_0 = x,$$

where the superscript ^{u} denotes that we are considering the system state with the control strategy u in operation. Here b and σ are functions $b : [0, \infty[\times \mathbb{R}^n \times \mathbb{U} \to \mathbb{R}^n$ and $\sigma : [0, \infty[\times \mathbb{R}^n \times \mathbb{U} \to \mathbb{R}^{n\times m}$, where $\mathbb{U}$ is the *control set* (the set of values that the control input can take). Often we will choose $\mathbb{U} = \mathbb{R}^q$, but this is not necessary.

Definition 6.1.1. The control strategy $u = \{u_t\}$ is called an *admissible strategy* if

1. u_t is an $\mathcal{F}_t$-adapted stochastic process; and

2. $u_t(\omega) \in \mathbb{U}$ for every (ω, t); and
3. the equation for X_t^u has a unique solution.

Remark 6.1.2. We will always consider the Wiener process W_t to be fixed, and require that X_t^u has a strong solution for admissible u. In a more general theory, it is often not clear whether strong solutions exist (e.g., for bang-bang controls), and such a definition may be too restrictive; it is not uncommon to require admissible u only to define a weak solution. (See chapter 5 for comments on weak vs. strong solutions).

There is a special type of control strategy that will be particularly important.

Definition 6.1.3. An admissible strategy u is called a *Markov strategy* if it is of the form $u_t = \alpha(t, X_t^u)$ for some function $\alpha : [0, \infty[\times \mathbb{R}^n \to \mathbb{U}$.

The reason for this terminology is clear: for a Markov strategy, the system state X_t^u is a Markov process (this is not true in general, where the control u_t may depend on the entire past history—it is only required to be $\mathcal{F}_t$-measurable!) Such strategies are important for two reasons: first, a Markov strategy is much easier to implement in practice than an arbitrary control functional; and second, we will find that the methods developed in this chapter automatically give rise to Markov strategies.

The goal of a control engineer is to design an admissible control strategy u to achieve a particular purpose. The design process, methods and machinery will obviously depend heavily on how we formulate the control goal. We already encountered one type of control goal in example 5.5.7: the goal was to find a controller u_t which would make an equilibrium point of the controlled SDE globally stable. The control goals which we will consider in this chapter and in the following chapters are of a rather different type; we are concerned here with *optimal control*. To this end, we will introduce a suitable *cost functional* that attaches to each admissible control strategy u a cost $J[u]$; the idea is to penalize undesirable behavior by giving it a large cost, while desirable behavior is encouraged by attaching to it a low cost. The goal is then to find, if possible, an *optimal* strategy u^* which minimizes this cost functional.

In this chapter we will investigate three common types of cost functionals:

1. For optimal control on the *finite time horizon* $[0, T]$, we introduce

$$J[u] = \mathbb{E}\left[\int_0^T w(s, X_s^u, u_s)\, ds + z(X_T^u)\right],$$

where $w : [0, T] \times \mathbb{R}^n \times \mathbb{U} \to \mathbb{R}$ (the *running cost*) and $z : \mathbb{R}^n \to \mathbb{R}$ (the *terminal cost*) are measurable functions and $T < \infty$ is the *terminal time*.

2. On an *indefinite time horizon*, we set

$$J[u] = \mathbb{E}\left[\int_0^{\tau^u} w(X_s^u, u_s)\, ds + z(X_{\tau^u}^u)\right],$$

where $w : S \times \mathbb{U} \to \mathbb{R}$ and $z : \partial S \to \mathbb{R}$ are measurable functions and the stopping time τ^u is the first exit time of X_t^u from $S \subset \mathbb{R}^n$ (with boundary ∂S).

3. On an *infinite time horizon*, we use either the discounted cost criterion

$$J_\lambda[u] = \mathbb{E}\left[\int_0^\infty e^{-\lambda s}\, w(X_s^u, u_s)\, ds\right],$$

or we can use a time-average cost criterion of the form

$$J[u] = \limsup_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\int_0^T w(X_s^u, u_s)\, ds\right],$$

where $w : \mathbb{R}^n \times \mathbb{U} \to \mathbb{R}$ is measurable.

The various types of cost functionals are not so dissimilar; once we figure out how to solve one of them, we can develop the other ones without too much trouble.

Remark 6.1.4. These are the most common types of cost functionals found in applications; we have seen some examples in the Introduction, and we will encounter more examples throughout this chapter. Others control costs have been considered as well, however (particularly the risk-sensitive cost criteria); see, e.g., [Bor05] for an overview of the various cost structures considered in the literature.

To motivate the development in the following sections, let us perform an illuminating but heuristic calculation; in particular, we will introduce nontrivial assumptions left and right and throw caution to the wind for the time being. What we will gain from this is a good intuition on the structure of the problem, armed with which we can proceed to obtain some genuinely useful results in the following sections.

The dynamic programming principle

In the remainder of this section, we will concentrate on the finite time horizon case, and we simplify life by restricting attention to Markov controls only. Fix the control problem (choose b, σ, w, z, T), and note that for any admissible Markov strategy u

$$\mathbb{E}\left[\int_t^T w(s, X_s^u, u_s)\, ds + z(X_T^u)\,\middle|\, \mathcal{F}_t\right] = \mathbb{E}\left[\int_t^T w(s, X_s^u, u_s)\, ds + z(X_T^u)\,\middle|\, X_t^u\right] \equiv J_t^u(X_t^u),$$

for $t \in [0,T]$, where we have used the Markov property of X_t^u (as u is a Markov strategy). The measurable function $J_t^u(x)$ is called the *cost-to-go* of the strategy u. You can interpret $J_t^u(x)$ as the portion of the total cost of the strategy u incurred in the time interval $[t,T]$, given that the control strategy in operation on the time interval $[0,t]$ has left us in the state $X_t^u = x$. In particular, $J_0^u(x)$ is the total cost of the strategy u if we start our system in the non-random state $X_0 = x$.

Remark 6.1.5. As we have defined a Markov process as a process X_t that satisfies $\mathbb{E}(f(X_t)|\mathcal{F}_s) = \mathbb{E}(f(X_t)|X_s)$ for bounded measurable f, the equality above may not be entirely obvious. The expression does follow from the following fundamental fact.

Lemma 6.1.6. *If X_t is an $\mathcal{F}_t$-Markov process, then $\mathbb{E}(F|\mathcal{F}_s) = \mathbb{E}(F|X_s)$ for any $\sigma\{X_t : t \in [s,\infty[\}$-measurable random variable F.*

Proof. First, note that for $t > r \geq s$ and bounded measurable f, g, $\mathbb{E}(f(X_t)g(X_r)|\mathcal{F}_s) = \mathbb{E}(\mathbb{E}(f(X_t)|\mathcal{F}_r)g(X_r)|\mathcal{F}_s) = \mathbb{E}(\mathbb{E}(f(X_t)|X_r)g(X_r)|\mathcal{F}_s) = \mathbb{E}(\mathbb{E}(f(X_t)|X_r)g(X_r)|X_s)$, using the Markov property and the fact that $\mathbb{E}(f(X_t)|X_r)g(X_r)$ is a bounded measurable function of X_r. By induction, $\mathbb{E}(f_1(X_{t_1})\cdots f_n(X_{t_n})|\mathcal{F}_s) = \mathbb{E}(f_1(X_{t_1})\cdots f_n(X_{t_n})|X_s)$ for any $n < \infty$, bounded measurable $f_1, \ldots, f_n$, and times $t_1, \ldots, t_n \geq s$.

Next, using the classical Stone-Weierstrass theorem, we find that any continuous function of n variables with compact support can be approximated uniformly by linear combinations of functions of the form $f_1(x_1)\cdots f_n(x_n)$, where f_i are continuous functions with compact support. Hence using dominated convergence, we find that $\mathbb{E}(f(X_{t_1},\ldots,X_{t_n})|\mathcal{F}_s) = \mathbb{E}(f(X_{t_1},\ldots,X_{t_n})|X_s)$ for any continuous f with compact support.

Finally, successive approximation establishes the claim for every $\sigma\{X_t : t \in [s,\infty[\}$-measurable random variable F. This follows exactly as in the proof of lemma 4.6.3. □

We would like to find an *optimal* control strategy u^*. Throughout this section we will assume that such a strategy exists, at least within the class of Markov strategies. In fact, for the purpose of this section, let us make a more daring assumption: *we assume that there exists an admissible Markov strategy u^* which satisfies $J_t^{u^*}(x) \leq J_t^u(x)$ for every admissible Markov strategy u, for all $t \in [0,T]$ and for all x.* This is certainly not always justified! However, let us go with it for the time being. Given the existence of this strategy u^*, we would like to find a way to actually *compute* what the strategy is. It is not at all obvious how to do this: minimizing directly over all admissible Markov strategies is hardly a feasible technique, even when significant computational resources are available! Instead, we will attempt to simplify matters by splitting up the optimization problem into a collection of smaller optimization problems.

The idea behind the methods in this chapter is the well known *dynamic programming principle* due to Bellman. The premise of this method is that it is not necessary to optimize the control strategy u over the entire time interval $[0,T]$ at once: we can divide the time interval into smaller chunks, and optimize over each individually. To this end, let us introduce the *value function* $V_t(x) = J_t^{u^*}(x)$; this is the optimal cost-to-go over the interval $[t,T]$. We claim that $V_t(x)$ satisfies the recursion

$$V_r(X_r^u) = \min_{u'} \mathbb{E}\left[\int_r^t w(s, X_s^{u'}, u'_s)\,ds + V_t(X_t^{u'})\,\middle|\, X_r^{u'}\right], \qquad 0 \leq r \leq t \leq T,$$

where the minimum is taken over all admissible Markov strategies u' that coincide with u on the interval $[0,r[$, and that this minimum is attained by the strategy which coincides with the optimal strategy u^* on the inteval $[r,t]$. Before we establish this claim, let us see why this is useful. Split the interval $[0,T]$ up into chunks $[0,t_1]$, $[t_1,t_2]$, …, $[t_n,T]$. Clearly $V_T(x) = z(x)$. We can now obtain $V_{t_n}(x)$ by computing the minimum above with $r = t_n$ and $t = T$, and this immediately gives us the optimal strategy on the interval $[t_n,T]$. Next, we can compute the optimal strategy on the previous interval $[t_{n-1},t_n]$ by minimizing the above expression with $r = t_{n-1}$, $t = t_n$ (as we now know $V_{t_n}(x)$ from the previous minimization), and iterating this procedure gives the optimal strategy u^* on the entire interval $[0,T]$. We will see below that this idea becomes particularly powerful if we let the partition size go to

zero: the calculation of the optimal control then becomes a *pointwise* minimization (i.e., separately for every time t), which is particularly straightforward to compute!

Let us now justify the dynamic programming principle. We begin by establishing a recursion for the cost-to-go: for any admissible Markov strategy u, we have

$$J_r^u(X_r^u) = \mathbb{E}\left[\int_r^t w(s, X_s^u, u_s)\,ds + J_t^u(X_t^u)\,\middle|\, X_r^u\right], \qquad 0 \le r \le t \le T.$$

This follows immediately from the definition of $J_r^u(x)$, using the Markov property and the tower property of the conditional expectation. Now choose u' to be a strategy that coincides with u on the interval $[0, t[$, and with u^* on the interval $[t, T]$. Then

$$V_r(X_r^u) \le J_r^{u'}(X_r^{u'}) = \mathbb{E}\left[\int_r^t w(s, X_s^{u'}, u'_s)\,ds + V_t(X_t^{u'})\,\middle|\, X_r^{u'}\right],$$

where we have used that $V_r(x) \le J_r^u(x)$ for any admissible Markov strategy u (by assumption), that X_s^u only depends on the strategy u in the time interval $[0, s[$, and that $J_s^u(x)$ only depends on u in the interval $[s, T]$ (use the Markov property). On the other hand, if we choose u' such that it coincides with u^* in the interval $[r, T]$, then we obtain this expression with equality rather than inequality using precisely the same reasoning. The dynamic programming recursion follows directly.

Remark 6.1.7 (Martingale dynamic programming principle). There is an equivalent, but more probabilistic, point of view on the dynamic programming principle which is worth mentioning (it will not be used in the following). Define the process

$$M_t^u = \int_0^t w(s, X_s^u, u_s)\,ds + V_t(X_t^u)$$

for every admissible Markov strategy u. You can easily establish (using the Markov property) that the dynamic programming principle is equivalent to the following statement: M_t^u is always a submartingale, while it is a martingale for $u = u^*$.

The Bellman equation

To turn the dynamic programming principle into a useful method, let us introduce some more assumptions (just go along with this for the time being!). Suppose that $V_t(x)$ is C^1 in t and C^2 in x; then, using Itô's rule,

$$V_t(X_t^u) = V_r(X_r^u) + \int_r^t \left\{\frac{\partial V_s}{\partial s}(X_s^u) + \mathscr{L}_s^u V_s(X_s^u)\right\} ds + \text{local martingale},$$

where $\mathscr{L}$ is the generator of the stochastic differential equation with the admissible Markov control u in operation, defined as in proposition 5.2.4 (note that it may depend on time in the current setting). If we additionally assume that the local martingale is in fact a martingale, then we obtain after some rearranging

$$V_r(X_r^u) = \mathbb{E}\left[\int_r^t \left\{-\frac{\partial V_s}{\partial s}(X_s^u) - \mathscr{L}_s^u V_s(X_s^u)\right\} ds + V_t(X_t^u)\,\middle|\, X_r^u\right].$$

But then we conclude, using the dynamic programming principle, that

$$\mathbb{E}\left[\int_r^t \left\{\frac{\partial V_s}{\partial s}(X_s^u) + \mathscr{L}_s^u V_s(X_s^u) + w(s, X_s^u, u_s)\right\} ds \,\middle|\, X_r^u\right] \geq 0$$

for every $0 \leq r \leq t \leq T$, and moreover the inequality becomes an equality if u coincides with u^* on the interval $[r, T]$. Thus, at the very least formally (evaluate the derivative with respect to t at $t = r$), we obtain the equation

$$\min_u \left\{\frac{\partial V_s}{\partial s}(X_s^u) + \mathscr{L}_s^u V_s(X_s^u) + w(s, X_s^u, u_s)\right\} = 0,$$

or, using the pathwise nature of this equation,

$$\min_{\alpha \in \mathbb{U}} \left\{\frac{\partial V_s(x)}{\partial s} + \mathscr{L}_s^\alpha V_s(x) + w(s, x, \alpha)\right\} = 0.$$

This is called the *Bellman equation*, and is "merely" an (extremely) nonlinear PDE.

Remark 6.1.8. To write the equation in more conventional PDE notation, note that we can write $\mathscr{L}_s^\alpha V_s(x) + w(s, x, \alpha)$ as a function H' of α, s, x and the first and second derivatives of $V_s(x)$. Hence the minimum of H' over α is simply some (highly nonlinear) function $H(s, x, \partial V_s(x), \partial^2 V_s(x))$, and the Bellman equation reads

$$\frac{\partial V_s(x)}{\partial s} + H(s, x, \partial V_s(x), \partial^2 V_s(x)) = 0.$$

We will encounter specific examples later on where this PDE can be solved explicitly.

If we can find a solution to the Bellman equation (with the terminal condition $V_T(x) = z(x)$) then we should be done: after all, the minimum over α (which depends both on s and x) must coincide with the optimal Markov control $u_t^* = \alpha(t, X_t^{u^*})$. Note that what we have done here is precisely the limit of the recursive procedure described above when the partition size goes to zero: we have reduced the computation to a *pointwise* optimization for every time s separately; indeed, the minimum above is merely over the set $\mathbb{U}$, not over the set of $\mathbb{U}$-valued control strategies on $[0, T]$. This makes finding optimal control strategies, if not easy, at least computationally feasible.

How to proceed?

The previous discussion is only intended as motivation. We have made various entirely unfounded assumptions, *which you should immediately discard from this point onward*. Let us take a moment for orientation; where can one proceed from here?

One direction in which we could go is the development of the story we have just told "for real", replacing all our assumptions by actual mathematical arguments. The assumption that an optimal control strategy exists and the obsession with Markov strategies can be dropped: in fact, one can show that the dynamic programming principle always holds (under suitable technical conditions, of course), regardless of whether an optimal strategy exists, provided we replace all the minima by infima! In

other words, the *infimum* of the cost-to-go always satisfies a recursion in the form encountered above. Moreover, we can drop the assumption that the value function is sufficiently smooth, and the Bellman equation will still hold under surprisingly general conditions—provided that we introduce an appropriate theory of weak solutions. The highly fine-tuned theory of *viscosity solutions* is designed especially for this purpose, and provides "just the right stuff" to build the foundations of a complete mathematical theory of optimal stochastic control. This direction is highly technical, however, while the practical payoff is not great: though there are applications of this theory, in particular in the analysis of numerical algorithms and in the search for near-optimal controls (which might be the only recourse if optimal controls do not exist), the main results of this theory are much more fundamental than practical in nature.

We will take the perpendicular direction by turning the story above upside down. Rather than starting with the optimal control problem, and showing that the Bellman equation follows, we will start with the Bellman equation (regarded simply as a nonlinear PDE) and suppose that we have found a solution. We will then show that this solution does indeed coincide with the value function of an optimal control problem, and that the control strategy obtained from the minimum in the Bellman equation is indeed optimal. This procedure is called *verification*, and is extremely practical: it says that if we can actually find a nice solution to the Bellman equation, then that solution gives an optimal control, which is what we care about in practice. This will allow us to solve a variety of control problems, while avoiding almost all technicalities.

Note that we previously encountered a similar tradeoff between the direct approach and verification: our discussion of the Kolmogorov backward equation is of the verification type. See remark 5.2.7 for further discussion on this matter.

Remark 6.1.9. It should be noted that stochastic optimal control problems are much better behaved, in general, than their deterministic counterparts. In particular, hardly any deterministic optimal control problem admits a "nice" solution to the Bellman equation, so that the approach of this chapter would be very restrictive in the deterministic case; however, the noise in our equations actually regularizes the Bellman equation somewhat, so that sufficiently smooth solutions are not uncommon (results in this direction usually follow from the theory of parabolic PDEs, and need not have much probabilistic content). Such regularity issues are beyond our scope, but see [FR75, section VI.6] and [FS06, section IV.4] for some details and further references.

Before we move on, let us give a simple example where the optimal control does not exist. This is very common, particularly if one is not careful in selecting a suitable cost functional, and it is important to realize the cause of such a problem.

Example 6.1.10. Consider the one-dimensional control system $dX_t^u = u_t\,dt + dW_t$, where our goal is to bring X_t^u as close as possible to zero by some terminal time T. It seems reasonable, then, to use a cost functional which only has a terminal cost: for example, consider the functional $J[u] = \mathbb{E}((X_T^u)^2)$. Using the Itô rule, we obtain

$$\mathbb{E}((X_T^u)^2) = \mathbb{E}((X_0)^2) + \int_0^T \mathbb{E}(2u_s X_s^u + 1)\,ds.$$

Now consider admissible Markov strategies of the form $u_t = -cX_t^u$, where $c > 0$ is some gain constant. Substituting into the previous expression, we find explicitly

$$\mathbb{E}((X_T^u)^2) = \frac{1}{2c} - e^{-2cT}\,\frac{1 - 2c\,\mathbb{E}((X_0)^2)}{2c}.$$

Evidently we can make the cost $J[u]$ arbitrarily close to zero by choosing a sufficiently large gain c. But $u_t = -\infty X_t^u$ is obviously not an admissible control strategy, and you can easily convince yourself that no admissible control strategy can achieve zero cost (as this would require the control to instantaneously set X_t^u to zero and keep it there). Hence an optimal control does not exist in this case. Similarly, the Bellman equation also fails to work here: we would like to write

$$\min_{\alpha\in\mathbb{R}}\left\{\frac{\partial V_s(x)}{\partial s} + \frac{1}{2}\frac{\partial^2 V_s(x)}{\partial x^2} + \alpha\,\frac{\partial V_s(x)}{\partial x}\right\} = 0,$$

but a minimum is clearly not attained (set $\alpha = -c\,\partial V_s(x)/\partial x$ with c arbitrarily large).

The problem is that we have not included a control-dependent term in the cost functional; the control is "free", and so we can apply an arbitrarily large gain without any negative consequences. In order to obtain a control problem which does have an optimal solution, we need to attach a large cost to control strategies that take large values. The easiest way to do this is to introduce a cost of the form

$$J[u] = \mathbb{E}\left[C\int_0^T (u_s)^2\,ds + (X_T^u)^2\right],$$

where the constant $C > 0$ adjusts the tradeoff between the magnitude of the control and the distance of the terminal state X_T^u from the origin. In this case, the Bellman equation does make sense: we obtain the *Hamilton-Jacobi PDE*

$$\begin{aligned} 0 &= \min_{\alpha\in\mathbb{R}}\left\{\frac{\partial V_s(x)}{\partial s} + \frac{1}{2}\frac{\partial^2 V_s(x)}{\partial x^2} + \alpha\,\frac{\partial V_s(x)}{\partial x} + C\alpha^2\right\} \\ &= \frac{\partial V_s(x)}{\partial s} + \frac{1}{2}\frac{\partial^2 V_s(x)}{\partial x^2} - \frac{1}{4C}\left(\frac{\partial V_s(x)}{\partial x}\right)^2, \end{aligned}$$

which has a smooth solution. The verification theorem in the next section then allows us to compute explicitly an optimal control strategy.

6.2 Verification: finite time horizon

Armed with our newly built intuition, we can start cranking out verification theorems. Compared to the somewhat complicated dynamic programming theory, the proofs of these simple results should seem particularly elegant!

In the current section, we work on a finite time horizon. Let us therefore fix

$$J[u] = \mathbb{E}\left[\int_0^T w(s, X_s^u, u_s)\,ds + z(X_T^u)\right].$$

We consider a controlled stochastic differential equation of the form

$$dX_t^u = b(t, X_t^u, u_t)\,dt + \sigma(t, X_t^u, u_t)\,dW_t,$$

and define the generator $\mathscr{L}_t^\alpha$, $\alpha \in \mathbb{U}$ as

$$\mathscr{L}_t^\alpha g(x) = \sum_{i=1}^n b^i(t,x,\alpha)\,\frac{\partial g}{\partial x^i}(x) + \frac{1}{2}\sum_{i,j=1}^n \sum_{k=1}^m \sigma^{ik}(t,x,\alpha)\sigma^{jk}(t,x,\alpha)\frac{\partial^2 g}{\partial x^i\,\partial x^j}(x).$$

We now have the following result.

Proposition 6.2.1. *Suppose there is a $V_t(x)$, which is C^1 in t and C^2 in x, such that*

$$\frac{\partial V_t(x)}{\partial t} + \min_{\alpha\in\mathbb{U}}\left\{\mathscr{L}_t^\alpha V_t(x) + w(t,x,\alpha)\right\} = 0, \qquad V_T(x) = z(x),$$

and $|\mathbb{E}(V_0(X_0))| < \infty$, and choose a minimum (which we implicitly assume to exist)

$$\alpha^*(t,x) \in \operatorname*{argmin}_{\alpha\in\mathbb{U}}\left\{\mathscr{L}_t^\alpha V_t(x) + w(t,x,\alpha)\right\}.$$

Denote by $\mathfrak{K}$ the class of admissible strategies u such that

$$\sum_{i=1}^n\sum_{k=1}^m \int_0^t \frac{\partial V_s}{\partial x^i}(X_s^u)\,\sigma^{ik}(s, X_s^u, u_s)\,dW_s^k$$

is a martingale (rather than a local martingale), and suppose that the control $u_t^* = \alpha^*(t, X_t^{u^*})$ defines an admissible Markov strategy which is in $\mathfrak{K}$. Then $J[u^*] \le J[u]$ for any $u \in \mathfrak{K}$, and $V_t(x) = J_t^{u^*}(x)$ is the value function for the control problem.

Remark 6.2.2. Note that $J[u^*] \le J[u]$ for any $u \in \mathfrak{K}$, i.e., u is not necessarily Markov (though the *optimal* strategy is always necessarily Markov if it is obtaind from a Bellman equation). On the other hand, we are restricted to admissible strategies which are sufficiently integrable to be in $\mathfrak{K}$; this is inevitable without some further hypotheses. It should be noted that such an integrability condition is often added to the definition of an admissible control strategy, i.e., we could interpret $\mathfrak{K}$ as the class of 'truly' admissible strategies. In applications, this is rarely restrictive.

Proof. For any $u \in \mathfrak{K}$, we obtain using Itô's rule and the martingale assumption

$$\mathbb{E}(V_0(X_0)) = \mathbb{E}\left[\int_0^T \left\{-\frac{\partial V_s}{\partial s}(X_s^u) - \mathscr{L}_s^{u_s} V_s(X_s^u) + V_T(X_T^u)\right\} ds\right].$$

But using $V_T(x) = z(x)$ and the Bellman equation, we find that

$$\mathbb{E}(V_0(X_0)) \le \mathbb{E}\left[\int_0^T w(s, X_s^u, u_s)\,ds + z(X_T^u)\right] = J[u].$$

On the other hand, if we set $u = u^*$, then we obtain $\mathbb{E}(V_0(X_0)) = J[u^*]$ following exactly the same steps. Hence $J[u^*] \le J[u]$ for all $u \in \mathfrak{K}$. The fact that $V_t(x) = J_t^{u^*}(x)$ follows easily in a similar manner (use Itô's rule and the martingale assumption), and the proof is complete. □

Let us show off this result with some interesting examples.

Example 6.2.3 (Tracking a particle under a microscope). In the Introduction, we discussed the example the problem of tracking a particle under a microscope in several different settings. We are finally in a position to start solving this problem. We proceed here in the simplest setting, and will return to this problem several times in this chapter and in the next chapter. Recall the the system was described by the pair of equations

$$\frac{dz_t}{dt} = \beta u_t, \qquad x_t = x_0 + \sigma W_t,$$

where z_t is the position of the slide relative to the focus of the microscope, x_t is the position of the particle we wish to view under the microscope relative to the center of the slide, $\beta \in \mathbb{R}$ is the gain in our servo loop and $\sigma > 0$ is the diffusion constant of the particle. We would like to keep the particle in focus, i.e., we would like to keep $x_t + z_t$ as close to zero as possible. However, we have to introduce a power constraint on the control as well, as we cannot drive the servo motor with arbitrarily large input powers. We thus introduce the control cost (see the Introduction)

$$J[u] = \mathbb{E}\left[\frac{p}{T}\int_0^T (x_t + z_t)^2\,dt + \frac{q}{T}\int_0^T (u_t)^2\,dt\right],$$

where $p, q > 0$ allow us to select the tradeoff between good tracking and low feedback power. To get rid of the pesky T^{-1} terms, let us define $P = p/T$ and $Q = q/T$.

As the control cost only depends on $x_t + z_t$, it is more convenient to proceed directly with this quantity. That is, define $e_t = x_t + z_t$, and note that

$$de_t = \beta u_t\,dt + \sigma\,dW_t, \qquad J[u] = \mathbb{E}\left[P\int_0^T (e_t)^2\,dt + Q\int_0^T (u_t)^2\,dt\right].$$

We obtain the Bellman equation

$$\begin{aligned} 0 &= \frac{\partial V_t(x)}{\partial t} + \min_{\alpha\in\mathbb{R}}\left\{\frac{\sigma^2}{2}\frac{\partial^2 V_t(x)}{\partial x^2} + \beta\alpha\,\frac{\partial V_t(x)}{\partial x} + Px^2 + Q\alpha^2\right\} \\ &= \frac{\partial V_t(x)}{\partial t} + \frac{\sigma^2}{2}\frac{\partial^2 V_t(x)}{\partial x^2} - \frac{\beta^2}{4Q}\left(\frac{\partial V_t(x)}{\partial x}\right)^2 + Px^2 \end{aligned}$$

with $V_T(x) = 0$ (as there is no terminal cost), and moreover

$$\alpha^*(t,x) = \operatorname*{argmin}_{\alpha\in\mathbb{R}}\left\{\beta\alpha\,\frac{\partial V_t(x)}{\partial x} + Q\alpha^2\right\} = -\frac{\beta}{2Q}\frac{\partial V_t(x)}{\partial x}.$$

We need to solve the Bellman equation. To this end, plug the following *ansatz* into the equation: $V_t(x) = a_t x^2 + b_t$. This gives, using $V_T(x) = 0$,

$$\frac{da_t}{dt} + P - \frac{\beta^2}{Q}\,a_t^2 = 0, \quad a_T = 0, \qquad \frac{db_t}{dt} + \sigma^2 a_t = 0, \quad b_T = 0.$$

With a little work, we can solve these equations explicitly:

$$a_t = \frac{\sqrt{PQ}}{\beta} \tanh\left(\beta\sqrt{\frac{P}{Q}}\,(T-t)\right), \quad b_t = \frac{Q\sigma^2}{\beta^2} \log\left(\cosh\left(\beta\sqrt{\frac{P}{Q}}\,(T-t)\right)\right).$$

Now note that $V_t(x)$ is smooth in x and t and that $\alpha^*(t,x)$ is uniformly Lipschitz on $[0,T]$. Hence if we assume that $\mathbb{E}((x_0+z_0)^2) < \infty$ (surely a reasonable requirement in this application!), then by theorem 5.1.3 we find that the feedback control

$$u_t^* = \alpha^*(t,e_t) = -\sqrt{\frac{P}{Q}}\,\tanh\left(\beta\sqrt{\frac{P}{Q}}\,(T-t)\right)(x_t+z_t)$$

satisfies $u_t^* \in \mathfrak{K}$. Thus, by proposition 6.2.1, u_t^* is an optimal control strategy.

Example 6.2.4 (Optimal portfolio selection). The following example comes from finance. We consider a single stock with average return $\mu > 0$ and volatility $\sigma > 0$, and a bank account with interest rate $r > 0$. This means that if we invest one dollar in stock or in the bank, respectively, at time zero, then at any later time t our bank account will contain R_t dollars and we will own S_t dollars worth of stock, where

$$dS_t = \mu S_t\,dt + \sigma S_t\,dW_t, \quad S_0 = 1, \qquad dR_t = rR_t\,dt, \quad R_0 = 1.$$

We now assume that we can modify our investment at any point in time. However, we only consider *self-financing* investment strategies: i.e., we begin with some starting capital $X_0 > 0$ (to be divided between the bank account and the stock), and we subsequently only transfer money between the bank account and the stock (without adding in any new money from the outside). Denote by X_t our total wealth at time t, and by u_t the fraction of our wealth that is invested in stock at time t (the remaining fraction $1-u_t$ being in the bank). Then the self-financing condition implies that

$$dX_t = \{\mu u_t + r(1-u_t)\}X_t\,dt + \sigma u_t X_t\,dW_t.$$

This can be justified as a limit of discrete time self-financing strategies; you have seen how this works in one of the homeworks, so we will not elaborate further.

Our goal is (obviously) to make money. Let us thus fix a terminal time T, and try to choose a strategy u_t that *maximizes* a suitable functional U of our total wealth at time T; in other words, we choose the cost functional $J[u] = \mathbb{E}(-U(X_T^u))$ (the minus sign appears as we have chosen, as a convention, to *minimize* our cost functionals). How to choose the *utility function* U is a bit of an art; the obvious choice $U(x) = x$ turns out not to admit an optimal control if we set $\mathbb{U} = \mathbb{R}$, while if we set $\mathbb{U} = [0,1]$ (we do not allow borrowing money or selling short) then we get a rather boring answer: we should always put all our money in stock if $\mu > r$, while if $\mu \le r$ we should put all our money in the bank (verify this using proposition 6.2.1!)

Other utility functions, however, can be used to encode our risk preferences. For example, suppose that U is nondecreasing and concave, e.g., $U(x) = \log(x)$ (the *Kelly criterion*). Then the relative penalty for ending up with a low total wealth is much heavier than for $U(x) = x$, so that the resulting strategy will be less risky

(concave utility functions lead to *risk-averse* strategies, while the utility $U(x) = x$ is called *risk-neutral*). As such, we would expect the Kelly criterion to tell us to put some money in the bank to reduce our risk! Let us see whether this is the case.[1]

The Bellman equation for the Kelly criterion reads (with $\mathbb{U} = \mathbb{R}$)

$$\begin{aligned} 0 &= \frac{\partial V_t(x)}{\partial t} + \min_{\alpha \in \mathbb{R}} \left\{ \frac{\sigma^2 \alpha^2 x^2}{2} \frac{\partial^2 V_t(x)}{\partial x^2} + (\mu\alpha + r(1-\alpha)) x \frac{\partial V_t(x)}{\partial x} \right\} \\ &= \frac{\partial V_t(x)}{\partial t} + rx \frac{\partial V_t(x)}{\partial x} - \frac{(\mu - r)^2}{2\sigma^2} \frac{(\partial V_t(x)/\partial x)^2}{\partial^2 V_t(x)/\partial x^2} \end{aligned}$$

where $V_T(x) = -\log(x)$, and moreover

$$\alpha^*(t,x) = -\frac{\mu - r}{\sigma^2} \frac{\partial V_t(x)/\partial x}{x\, \partial^2 V_t(x)/\partial x^2},$$

provided that $\partial^2 V_t(x)/\partial x^2 > 0$ *for all* $x > 0$ (otherwise a minimum does not exist!). Once we have solved for $V_t(x)$, we must remember to check this assumption.

These unsightly expressions seem more hopeless than they actually are. Fill in the *ansatz* $V_t(x) = -\log(x) + b_t$; then we obtain the simple ODE

$$\frac{db_t}{dt} - C = 0, \quad b_T = 0, \qquad C = r + \frac{(\mu - r)^2}{2\sigma^2}.$$

Thus evidently $V_t(x) = -\log(x) - C(T-t)$ solves the Bellman equation, and moreover this function is smooth on $x > 0$ and $\partial^2 V_t(x)/\partial x^2 > 0$ as required. Furthermore, the corresponding control is $\alpha^*(t,x) = (\mu - r)/\sigma^2$, which is as regular as it gets. By theorem 5.1.3 (and by the fact that our starting capital $X_0 > 0$ is non-random), the conditions of propostion 6.2.1 are met and we find that $u_t = (\mu - r)/\sigma^2$ is indeed the optimal control. Evidently the Kelly criterion tells us to put money in the bank, provided that $\mu - r < \sigma^2$. On the other hand, if $\mu - r$ is large, it is advantageous to borrow money from the bank to invest in stock (this is possible in the current setting as we have chosen $\mathbb{U} = \mathbb{R}$, rather than restricting to $\mathbb{U} = [0,1]$).

6.3 Verification: indefinite time horizon

In this section and the next, we restrict ourselves to time-homogeneous control systems, i.e., we will let b and σ be independent of time t. This is not a restriction: if we wish to add time dependence, we can simply increase the dimension of the state space by one and consider time to be one of the states of the system. However, our results will look a little cleaner without the explicit time dependence. As we will see, the resulting control strategies conveniently do not depend on time either.

We thus proceed with the control system

$$dX_t^u = b(X_t^u, u_t)\, dt + \sigma(X_t^u, u_t)\, dW_t,$$

[1] Note that $\log(x)$ is not C^2 on $\mathbb{R}$; however, as a self-financed wealth process is always positive, everything goes through as usual through localization (see the remark after the proof of Itô's rule).

and consider minimizing the cost functional

$$J[u] = \mathbb{E}\left[\int_0^{\tau^u} w(X_s^u, u_s)\, ds + z(X_{\tau^u}^u)\right].$$

Here $\tau^u = \inf\{t : X_t^u \notin S\}$ where $S \subset \mathbb{R}^n$ is some bounded domain, and $w : S \times \mathbb{U} \to \mathbb{R}$ and $z : \partial S \to \mathbb{R}$ are the running and terminal costs, respectively.

For example, an interesting class of such problems is obtained if we set $w = 1$ and $z = 0$; then the cost is simply $J[u] = \mathbb{E}(\tau^u)$, and the corresponding control problem seeks to minimize the mean exit time from the domain S. If $w = -1$, on the other hand, then we seek to postpone exiting the domain as long as possible (on average).

Proposition 6.3.1. *Assume that S has compact closure $\overline{S}$ and $X_0 \in S$ a.s. Suppose there is a function $V : \overline{S} \to \mathbb{R}$ that is C^2 on $\overline{S}$ and satisfies (∂S is the boundary of S)*

$$\min_{\alpha\in\mathbb{U}} \{\mathscr{L}^\alpha V(x) + w(x,\alpha)\} = 0, \quad x \in S, \qquad V(x) = z(x), \quad x \in \partial S.$$

Choose a minimum (which we have implicitly assumed to exist)

$$\alpha^*(x) \in \operatorname*{argmin}_{\alpha\in\mathbb{U}} \{\mathscr{L}^\alpha V(x) + w(x,\alpha)\}.$$

Denote by $\mathfrak{K}$ the class of admissible strategies u such that $\tau^u < \infty$ a.s. and

$$\mathbb{E}\left[\sum_{i=1}^n \sum_{k=1}^m \int_0^{\tau^u} \frac{\partial V}{\partial x^i}(X_s^u)\, \sigma^{ik}(X_s^u, u_s)\, dW_s^k\right] = 0.$$

If $u_t^* = \alpha^*(X_t^{u^*})$ defines an admissible Markov strategy in $\mathfrak{K}$, then $J[u^*] \le J[u]$ for any $u \in \mathfrak{K}$, and the optimal cost can be expressed as $\mathbb{E}(V(X_0)) = J[u^*]$.

Proof. Using a simple localization argument and the assumption on $u \in \mathfrak{K}$, Itô's rule gives

$$\mathbb{E}(V(X_{\tau^u}^u)) = \mathbb{E}(V(X_0)) + \mathbb{E}\left[\int_0^{\tau^u} \mathscr{L}^{u_s} V(X_s^u)\, ds\right].$$

Using the Bellman equation and $X_{\tau^u}^u \in \partial S$, we obtain

$$\mathbb{E}(V(X_0)) \le \mathbb{E}\left[\int_0^{\tau^u} w(X_s^u, u_s)\, ds + z(X_{\tau^u}^u)\right] = J[u].$$

On the other hand, we obtain equality if $u = u^*$, so we are done. □

Example 6.3.2 (Tracking under a microscope II). We consider again the problem of tracking a particle under a microscope, but with a slightly different premise. Most microscopes have a field of view whose shape is a disc of some radius r around the focal point of the microscope. In other words, we will see the particle if it is within a distance r of the focus of the microscope, but we will have no idea where the particle is if it is outside the field of view. Given that we begin with the particle inside the field of view, our goal should thus be to keep the particle in the field of view as long

as possible by moving around the slide; once we lose the particle, we might as well give up. On the other hand, as before, we do not allow arbitrary controls: we have to impose some sort of power constraint to keep the feedback signal sane.

Let us study the following cost. Set $S = \{x : |x| < r\}$, let $\tau^u = \inf\{t : e_t^u \notin S\}$ (recall that $e_t = x_t + z_t$ is the position of the particle relative to the focus), and define

$$J[u] = \mathbb{E}\left[p\int_0^{\tau^u} (u_s)^2\,ds - q\,\tau^u\right] = \mathbb{E}\left[\int_0^{\tau^u} \{p\,(u_s)^2 - q\}\,ds\right]$$

where $p > 0$ and $q > 0$ are constants. We assume that $e_0 \in S$ a.s. A control strategy that minimizes $J[u]$ then attempts to make τ^u large (i.e., the time until we lose the particle is large), while keeping the total feedback power relatively small; the tradeoff between these conflicting goals can be selected by playing around with p and q.

To find the optimal strategy, we try to solve the Bellman equation as usual:

$$\begin{aligned} 0 &= \min_{\alpha\in\mathbb{R}}\left\{\frac{\sigma^2}{2}\frac{\partial^2 V(x)}{\partial x^2} + \beta\alpha\frac{\partial V(x)}{\partial x} + p\,\alpha^2 - q\right\} \\ &= \frac{\sigma^2}{2}\frac{\partial^2 V(x)}{\partial x^2} - q - \frac{\beta^2}{4p}\left(\frac{\partial V(x)}{\partial x}\right)^2 \end{aligned}$$

with the boundary conditions $V(r) = V(-r) = 0$, and a minimum is attained at

$$\alpha^*(x) = -\frac{\beta}{2p}\frac{\partial V(x)}{\partial x}.$$

But we can now solve the Bellman equation explicitly: it evidently reduces to a one-dimensional ODE for $\partial V(x)/\partial x$. Some work gives the solution

$$V(x) = \frac{2p\sigma^2}{\beta^2}\left[\log\left(\cos\left(\frac{r\beta\sqrt{q}}{\sigma^2\sqrt{p}}\right)\right) - \log\left(\cos\left(\frac{x\beta\sqrt{q}}{\sigma^2\sqrt{p}}\right)\right)\right],$$

while the minimum is attained at

$$\alpha^*(x) = -\sqrt{\frac{q}{p}}\,\tan\left(\frac{x\beta\sqrt{q}}{\sigma^2\sqrt{p}}\right),$$

provided that $r\beta\sqrt{q}/\sigma^2\sqrt{p}$ is sufficiently small; in fact, we clearly need to require $2r\beta\sqrt{q} < \pi\sigma^2\sqrt{p}$, as only in this case are $V(x)$ and $\alpha^*(x)$ in C^2 on $[-r, r]$. Apparently this magic inequality, which balances the various parameters in our control problem, determines whether an optimal control exists; you would have probably had a difficult time guessing this fact without performing the calculation!

It remains to verify the technical conditions of proposition 6.3.1, i.e., that the control strategy $u_t^* = \alpha^*(e_t)$ satisfies $\tau^{u^*} < \infty$ a.s. and the condition on the stochastic integral (clearly u_t^* is admissible, as $\alpha^*(x)$ is Lipschitz continuous on $[-r, r]$). The finiteness of $\mathbb{E}(\tau^{u^*})$ follows from lemma 6.3.3 below, while the stochastic integral condition follows from lemma 6.3.4 below. Hence u_t^* is indeed an optimal strategy.

The technical conditions of proposition 6.3.1 are not entirely trivial to check; the following two lemmas are often helpful in this regard, and can save a lot of effort.

Lemma 6.3.3. *Let X_t be the solution of the SDE $dX_t = b(X_t)\,dt + \sigma(X_t)\,dW_t$, where b and σ are assumed to be Lipschitz as usual, and suppose that $X_0 \in S$ a.s. for some bounded domain $S \subset \mathbb{R}^n$. If σ satisfies the nondegeneracy condition on S*

$$\sum_{i,j=1}^{n}\sum_{k=1}^{m} v^i \sigma^{ik}(x)\sigma^{jk}(x)v^j \geq \gamma \|v\|^2 \qquad \forall\, v \in \mathbb{R}^m,\ x \in S,$$

for some constant $\gamma > 0$, then $\tau_S = \inf\{t : X_t \notin S\}$ satisfies $\mathbb{E}(\tau_S) < \infty$.

Proof. Define the function $W(x) = k - (x^1 + \beta)^{2n}$, and calculate

$$\mathscr{L}W(x) = -2n\, b^1(x)\,(x^1+\beta)^{2n-1} - n(2n-1)\,(x^1+\beta)^{2n-2}\sum_{k=1}^{m}(\sigma^{1k}(x))^2.$$

Here k, β and n are suitable constants which we will currently choose. As S is bounded, we can choose $\beta \in \mathbb{R}$ such that $0 < c_1 < |x^1 + \beta| < c_2 < \infty$ for all $x \in S$. Next, note that as b is continuous on $\mathbb{R}^n$ it must be bounded on S; in particular, $|b^1(x)| < b_0$ for some $b_0 \in \mathbb{R}$ and all $x \in S$. Hence we can estimate, using the nondegeneracy condition,

$$\mathscr{L}W(x) < \{2nb_0c_2 - n(2n-1)\gamma/2\}(x^1+\beta)^{2n-2} \qquad \forall\, x \in S.$$

Clearly we can choose n sufficiently large so that the prefactor is bounded from above by $-c_3$ for some $c_3 > 0$; then we obtain $\mathscr{L}W(x) < -c_3\, c_1^{2n-2} < 0$ for all $x \in S$. Finally, we can choose k sufficiently large so that $W(x)$ is nonnegative.

It remains to show that the existence of W implies $\mathbb{E}(\tau_S) < \infty$. To this end, write

$$W(X_{t\wedge\tau_S}) = W(X_0) = \int_0^{t\wedge\tau_S} \mathscr{L}W(X_r)\,dr + \text{martingale},$$

where the stochastic integral is a martingale (rather than a local martingale) as the integrand is bounded on S. Taking the expectation and using $\mathscr{L}W(x) \leq -c_4$ ($c_4 > 0$) for $x \in S$, we find

$$\mathbb{E}(W(X_{t\wedge\tau_S})) \leq \mathbb{E}(W(X_0)) - c_4\,\mathbb{E}(t\wedge\tau_S).$$

But W is bounded on S, so we have established that $\mathbb{E}(t\wedge\tau_S) \leq K$ for some $K < \infty$ and for all t. Letting $t \to \infty$ and using monotone convergence establishes the result. □

Lemma 6.3.4. *Let τ be a stopping time such that $\mathbb{E}(\tau) < \infty$, and let u_t be an adapted process that satisfies $|u_t| \leq K$ for all $t \leq \tau$ and a $K < \infty$. Then $\mathbb{E}[\int_0^\tau u_s\,dW_s] = 0$.*

Proof. Define the stochastic process

$$M_t = \int_0^{t\wedge\tau} u_s\,dW_s.$$

As $\tau < \infty$ a.s., $M_t \to M_\infty$ as $t \to \infty$. We need to show that $\mathbb{E}(M_\infty) = 0$. To this end, note first that M_t is a martingale (not a local martingale), as u_s is bounded for $s \leq \tau$. Hence $\mathbb{E}(M_t) = 0$ for all $t < \infty$. We will show that $M_n \to M_\infty$ in $\mathcal{L}^2(\mathbb{P})$ (where $n \in \mathbb{N}$), from which the claim follows directly. To establish convergence in $\mathcal{L}^2(\mathbb{P})$, compute

$$\mathbb{E}((M_n - M_m)^2) = \mathbb{E}((M_n)^2) - \mathbb{E}((M_m)^2) = \mathbb{E}\left[\int_{m\wedge\tau}^{n\wedge\tau}(u_r)^2 dr\right] \leq K^2\,\mathbb{E}(n\wedge\tau - m\wedge\tau),$$

which converges to zero as $m, n \to \infty$ by dominated convergence (use that $\mathbb{E}(\tau) < \infty$). Hence M_n is a Cauchy sequence in $\mathcal{L}^2(\mathbb{P})$, and thus converges in $\mathcal{L}^2(\mathbb{P})$. We are done. □

6.4 Verification: infinite time horizon

We now proceed to the infinite time horizon. Little changes here, except that we have to be careful to define a meaningful cost functional. For example, in our tracking example on a finite time horizon, we cannot simply set the terminal time $T = \infty$; if we do that, then any control strategy will have infinite cost (why?). We can avoid this problem by adding a discounting term in the cost functional, as follows:

$$J_\lambda[u] = \mathbb{E}\left[\int_0^\infty e^{-\lambda s}\, w(X_s^u, u_s)\, ds\right].$$

Here $\lambda > 0$ is the *discounting factor*. Such a cost often makes sense in economic applications, where discounting is a natural thing to do (inflation will make one dollar at time t be worth much less than one dollar at time zero). Now if w is bounded, or if X_s^u does not grow too fast, then this cost is guaranteed to be finite and we can attempt to find optimal controls as usual. Alternatively, we can average over time by setting

$$J[u] = \limsup_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\int_0^T w(X_s^u, u_s)\, ds\right],$$

which might make more sense in applications which ought to perform well uniformly in time. Once again, if w does not grow too fast, this cost will be bounded.

Remark 6.4.1. It should be emphasized that these cost functionals, as well as those discussed in the previous sections, certainly do not exhaust the possibilities! There are many variations on this theme, and with your current intuition you should not have too much trouble obtaining related verification theorems. For example, try to work out a verification theorem for a discounted version of the indefinite time interval problem.

Let us now develop appropriate verification theorems for the costs $J_\lambda[u]$ and $J[u]$.

Proposition 6.4.2 (Discounted case). *Assume that $w(x,\alpha)$ is either bounded from below or from above. Suppose there is a $V(x)$ in C^2 such that $|\mathbb{E}(V(X_0))| < \infty$ and*

$$\min_{\alpha\in\mathbb{U}} \{\mathscr{L}^\alpha V(x) - \lambda V(x) + w(x,\alpha)\} = 0,$$

and choose a minimum (which we implicitly assume to exist)

$$\alpha^*(x) \in \operatorname*{argmin}_{\alpha\in\mathbb{U}} \{\mathscr{L}^\alpha V(x) - \lambda V(x) + w(x,\alpha)\}.$$

Denote by $\mathfrak{K}$ the admissible strategies u such that $e^{-\lambda t}\,\mathbb{E}(V(X_t^u)) \xrightarrow{t\to\infty} 0$ and

$$\sum_{i=1}^n \sum_{k=1}^m \int_0^t e^{-\lambda s}\, \frac{\partial V}{\partial x^i}(X_s^u)\, \sigma^{ik}(X_s^u, u_s)\, dW_s^k$$

is a martingale (rather than a local martingale), and suppose that the control $u_t^* = \alpha^*(X_t^{u^*})$ defines an admissible Markov strategy which is in $\mathfrak{K}$. Then $J_\lambda[u^*] \le J_\lambda[u]$ for any $u \in \mathfrak{K}$, and the optimal cost can be written as $\mathbb{E}(V(X_0)) = J_\lambda[u^*]$.

Proof. Applying Itô's rule to $V(X_t^u)\, e^{-\lambda t}$ and using the assumptions on $u \in \mathfrak{K}$,

$$\mathbb{E}(V(X_0)) - e^{-\lambda t}\, \mathbb{E}(V(X_t^u)) = \mathbb{E}\left[\int_0^t e^{-\lambda s}\{-\mathscr{L}^{u_s} V(X_s^u) + \lambda\, V(X_s^u)\}\, ds\right].$$

Using the Bellman equation, we find that

$$\mathbb{E}(V(X_0)) - e^{-\lambda t}\, \mathbb{E}(V(X_t^u)) \le \mathbb{E}\left[\int_0^t e^{-\lambda s}\, w(X_s^u, u_s)\, ds\right].$$

We may assume without loss of generality that w is either nonnegative or nonpositive; otherwise this is easily arranged by shifting the cost. Letting $t \to \infty$ using monotone convergence,

$$\mathbb{E}(V(X_0)) \le \mathbb{E}\left[\int_0^\infty e^{-\lambda s}\, w(X_s^u, u_s)\, ds\right] = J_\lambda[u].$$

But we obtain equality if we use $u = u^*$, so we are done. □

The time-average problem has a new ingredient: the function $V(x)$ no longer determines the optimal cost (note that on the infinite time horizon, the optimal cost is independent of X_0; on the other hand, the control must depend on x!). We need to introduce another free parameter for the Bellman equation to admit a solution.

Proposition 6.4.3 (Time-average case). *Suppose that $V(x)$ in C^2 and $\eta \in \mathbb{R}$ satisfy*

$$\min_{\alpha \in \mathbb{U}} \{\mathscr{L}^\alpha V(x) + w(x, \alpha) - \eta\} = 0,$$

and choose a minimum (which we implicitly assume to exist)

$$\alpha^*(x) \in \operatorname*{argmin}_{\alpha \in \mathbb{U}} \{\mathscr{L}^\alpha V(x) + w(x, \alpha) - \eta\}.$$

Denote by $\mathfrak{K}$ the class of admissible strategies u such that

$$\limsup_{T \to \infty} \frac{\mathbb{E}(V(X_0) - V(X_T^u))}{T} = 0,$$

and such that additionally

$$\sum_{i=1}^n \sum_{k=1}^m \int_0^t \frac{\partial V}{\partial x^i}(X_s^u)\, \sigma^{ik}(X_s^u, u_s)\, dW_s^k$$

is a martingale (rather than a local martingale), and suppose that the control $u_t^* = \alpha^*(X_t^{u^*})$ defines an admissible Markov strategy which is in $\mathfrak{K}$. Then $J[u^*] \le J[u]$ for any $u \in \mathfrak{K}$, and the optimal cost is given by $\eta = J[u^*]$.

Proof. Applying Itô's rule to $V(X_t^u)$ and using the assumptions on $u \in \mathfrak{K}$, we obtain

$$\frac{\mathbb{E}(V(X_0) - V(X_T^u))}{T} + \eta = \mathbb{E}\left[\frac{1}{T}\int_0^T \{\eta - \mathscr{L}^{u_s} V(X_s^u)\}\, ds\right] \le \mathbb{E}\left[\frac{1}{T}\int_0^T w(X_s^u, u_s)\, ds\right],$$

where we have already used the Bellman equation. Taking the limit gives

$$\eta \le \limsup_{T \to \infty}\, \mathbb{E}\left[\frac{1}{T}\int_0^T w(X_s^u, u_s)\, ds\right] = J[u].$$

But we obtain equality if we use $u = u^*$, so we are done. □

Some examples are in order.

Example 6.4.4 (Tracking under a microscope III). The goal is to repeat example 6.2.3 using discounted and time-average cost criteria. In particular, we consider

$$J_\lambda[u] = \mathbb{E}\left[p\int_0^\infty e^{-\lambda t}\,(x_t+z_t)^2\,dt + q\int_0^\infty e^{-\lambda t}\,(u_t)^2\,dt\right]$$

for the discounted cost, and we consider the time-average cost

$$J[u] = \limsup_{T\to\infty}\,\mathbb{E}\left[\frac{p}{T}\int_0^T (x_t+z_t)^2\,dt + \frac{q}{T}\int_0^T (u_t)^2\,dt\right].$$

Let us begin by investigating the discounted cost. The Bellman equation becomes

$$\begin{aligned}0 &= \min_{\alpha\in\mathbb{R}}\left\{\frac{\sigma^2}{2}\frac{\partial^2 V(x)}{\partial x^2} + \beta\alpha\,\frac{\partial V(x)}{\partial x} - \lambda\,V(x) + px^2 + q\alpha^2\right\}\\ &= \frac{\sigma^2}{2}\frac{\partial^2 V(x)}{\partial x^2} - \frac{\beta^2}{4q}\left(\frac{\partial V(x)}{\partial x}\right)^2 - \lambda\,V(x) + px^2,\end{aligned}$$

and, moreover, the minimal α is attained at

$$\alpha^*(x) = -\frac{\beta}{2q}\frac{\partial V(x)}{\partial x}.$$

To solve the Bellman equation, substitute the *ansatz* $V(x) = ax^2 + b$. We obtain

$$b = \frac{\sigma^2 a}{\lambda}, \qquad p - \lambda a - \frac{\beta^2 a^2}{q} = 0 \quad\Longrightarrow\quad a = -\frac{q\lambda \pm \sqrt{q^2\lambda^2 + 4pq\beta^2}}{2\beta^2}.$$

There are multiple solutions! Now what? The key is that every solution to the Bellman equation yields a candidate control $\alpha^*(x)$, but only one of these will satisfy the technical conditions in the verification. Let us check this. The candidate strategies are

$$\alpha_1^*(x) = \frac{\lambda + \sqrt{\lambda^2 + 4p\beta^2/q}}{2\beta}\,x, \qquad \alpha_2^*(x) = \frac{\lambda - \sqrt{\lambda^2 + 4p\beta^2/q}}{2\beta}\,x.$$

Note that $\alpha_1^*(x) = c_1 x$ with $\beta c_1 > \lambda$, while $\alpha_2^*(x) = -c_2 x$ with $\beta c_2 > 0$ (assuming $p > 0$; the case $p = 0$ is trivial, as then the optimal control is clearly $u_t = 0$). But

$$de_t = \beta c\,e_t\,dt + \sigma dW_t \quad\Longrightarrow\quad \frac{d}{dt}\,\mathbb{E}(V(e_t)) = 2\beta c\,\mathbb{E}(V(e_t)) - 2\beta bc + a\sigma^2.$$

Hence provided that $\mathbb{E}((e_0)^2) < \infty$, the quantity $\mathbb{E}(V(e_t))$ grows exponentially at a rate faster than λ for the control α_1^*, whereas $\mathbb{E}(V(e_t))$ is bounded for the control α_2^*. Hence α_2^* is the only remaining candidate control. It remains to check the martingale condition, but this follows immediately from theorem 5.1.3. Hence we conclude that $u_t^* = \alpha_2^*(e_t)$ is an optimal control for the discounted problem.

Let us now consider the time-average problem. The Bellman equation is

$$\begin{aligned} 0 &= \min_{\alpha\in\mathbb{R}} \left\{ \frac{\sigma^2}{2}\frac{\partial^2 V(x)}{\partial x^2} + \beta\alpha\,\frac{\partial V(x)}{\partial x} + px^2 + q\alpha^2 - \eta \right\} \\ &= \frac{\sigma^2}{2}\frac{\partial^2 V(x)}{\partial x^2} - \frac{\beta^2}{4q}\left(\frac{\partial V(x)}{\partial x}\right)^2 + px^2 - \eta, \end{aligned}$$

with the same minimal α as before. To solve the Bellman equation, substitute the *ansatz* $V(x) = ax^2$. We find that $\eta = \sigma^2 a$, while $a^2 = pq/\beta^2$. Once again there are two solutions, but repeating exactly the same arguments as in the discounted case shows that the only solution that is a viable candidate for being an optimal strategy is

$$\alpha^*(x) = -\sqrt{\frac{p}{q}}\,x.$$

Indeed, provided that $\mathbb{E}((e_0)^2) < \infty$, all the conditions are satisfied and we conclude that $u_t^* = \alpha^*(e_t)$ is an optimal time-average control strategy.

Remark 6.4.5. Note that the time-average optimal control strategy coincides with the limit of the finite time horizon optimal control as the terminal time $T \to \infty$, as well as with the limit of the discounted cost optimal control as the discounting factor $\lambda \to 0$. Heuristically, this is precisely what you would expect!

Remark 6.4.6. The previous example highlights that the solution to the Bellman equation need not be unique—if there are multiple solutions, the technical conditions of the verification theorem may tell us which one to choose! We will see this even more dramatically in the context of optimal stopping. On the other hand, you should realize that the optimal control strategy need not be unique; it is possible for there to be multiple optimal strategies, though they would have to have the same cost. There can also be no optimal strategies: we have seen plenty of examples of this already.

Example 6.4.7 (Tracking under a microscope III, cont.). In the Introduction, we considered studying the fundamental limitations of our tracking control system. In this context, it is of significant interest to compute the quantity

$$C(U) = \min_u \left\{ \limsup_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\int_0^T (e_t^u)^2\,dt\right] : \limsup_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\int_0^T u_t^2\,dt\right] \le U \right\},$$

which quantifies the best possible effectiveness of a tracking controller given a hard constraint on the average power in the feedback signal. Note that if we define

$$K(U) = \min_u \left\{ \limsup_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\int_0^T (e_t^u)^2\,dt\right] : \limsup_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\int_0^T u_t^2\,dt\right] = U \right\},$$

then $C(U) = \min_{U'\le U} K(U')$. Hence it suffices to compute the function $K(U)$.

How does one solve such a problem? The trick is to use the constant q in our previous cost functional as a Lagrange multiplier (we can set $p = 1$), i.e., we consider

$$J_{q,U}[u] = \limsup_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\int_0^T (e_t^u)^2\,dt + \frac{q}{T}\int_0^T u_t^2\,dt - qU\right].$$

Then we have $\min_u J_{q,U}[u] \le K(U)$ for all $q > 0$ (why?). Hence if we can find a $q > 0$ such that this inequality becomes an equality, then we have determined $K(U)$.

Let us work out the details. We already established above that

$$\min_u J_{q,U}[u] = \frac{\sigma^2\sqrt{q}}{\beta} - qU, \qquad \operatorname*{argmin}_u J_{q,U}[u] = u^* \quad \text{with} \quad u_t^* = -\frac{e_t}{\sqrt{q}}.$$

In particular, we can calculate explicitly (how?)

$$\limsup_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\int_0^T (e_t^{u^*})^2\,dt\right] = q \limsup_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\int_0^T (u_t^*)^2\,dt\right] = \frac{\sigma^2\sqrt{q}}{2\beta}.$$

Hence if we set $q = \sigma^4/4\beta^2U^2$, then

$$\limsup_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\int_0^T (u_t^*)^2\,dt\right] = U, \qquad \limsup_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\int_0^T (e_t^{u^*})^2\,dt\right] = \frac{\sigma^4}{4\beta^2 U},$$

but also $\min_u J_{q,U}[u] = \sigma^4/4\beta^2U$. Thus apparently

$$C(U) = \min_{U'\le U} K(U') = \min_{U'\le U} \frac{\sigma^4}{4\beta^2U'} = \frac{\sigma^4}{4\beta^2U}.$$

As expected, we can track better when we increase the gain β or the feedback power U, while we track worse if the particle has a larger diffusion constant σ.

6.5 The linear regulator

One of the most important classes of stochastic optimal control problems that admit explicit solution is the linear regulator problem. We have already seen a special case of this theory: two of the three tracking examples in the previous sections are simple examples of a linear regulator. The linear regulator is particularly important for applications in engineering and in the physical sciences (as opposed to financial applications which usually involve a somewhat different type of control problem; compare with the portfolio optimization example), and is remarkably ubiquitous. These results will become even more powerful in the next chapter, where they are combined with filtering theory, but for the time being we will set up this problem in a general setting under the assumption of complete observations (i.e., the control strategy is allowed to be an arbitrary admissible functional of the system state).

The linear regulator problem aims to control the stochastic differential equation

$$dX_t^u = A(t)X_t^u\,dt + B(t)u_t\,dt + C(t)\,dW_t,$$

where $A(t)$, $B(t)$ and $C(t)$ are time-dependent (but *non-random*) matrices of dimensions $n \times n$, $n \times k$, and $n \times m$, respectively, X_t^u is the n-dimensional system state (under the control strategy u), W_t is an m-dimensional Wiener process, and u_t is the k-dimensional control input. We consider the finite time horizon cost

$$J[u] = \mathbb{E}\left[\int_0^T \{(X_t^u)^*P(t)X_t^u + (u_t)^*Q(t)u_t\}\,dt + (X_T^u)^*R(X_T^u)\right],$$

where $T < \infty$ is the terminal time, $P(t)$ and $Q(t)$ are time-dependent (but non-random) $n \times n$ and $k \times k$ matrices that determine the state and control running cost, respectively, and R is a fixed (non-random) $n \times n$ matrix which determines the terminal cost. Let us make the following additional assumptions.

1. $\mathbb{E}(\|X_0\|^2) < \infty$;
2. $A(t), B(t), C(t), P(t), Q(t)$ are continuous on $t \in [0, T]$;
3. $P(t)$, $Q(t)$ and R are symmetric matrices (they can always be symmetrized);
4. $P(t)$ and R are positive semidefinite for all $t \in [0, T]$;
5. $Q(t)$ is positive definite on $t \in [0, T]$.

Our goal is to find a control strategy that minimizes $J[u]$.

Theorem 6.5.1 (Linear regulator, finite time). *Denote by $\{F(t)\}_{t\in[0,T]}$ the unique solution, with terminal condition $F(T) = R$, of the matrix Riccati equation*

$$\frac{d}{dt}F(t) + A(t)^*F(t) + F(t)A(t) - F(t)B(t)Q(t)^{-1}B(t)^*F(t) + P(t) = 0.$$

Then $u_t^* = -Q(t)^{-1}B(t)^*F(t)X_t^u$ is an optimal control for the cost $J[u]$.

Proof. We need to solve the Bellman equation. In the current setting, this is

$$0 = \frac{\partial V_t(x)}{\partial t} + \min_{\alpha\in\mathbb{R}^k}\left\{(A(t)x + B(t)\alpha)^*\nabla V_t(x) + \frac{1}{2}\nabla^* C(t)C(t)^*\nabla V_t(x) + x^*P(t)x + \alpha^*Q(t)\alpha\right\},$$

where we set $V_T(x) = x^*Rx$. As $Q(t)$ is positive definite, the minimum is attained at

$$\alpha^*(t,x) = -\frac{1}{2}Q(t)^{-1}B(t)^*\nabla V_t(x),$$

so the Bellman equation can be written as

$$0 = \frac{\partial V_t(x)}{\partial t} + \frac{1}{2}\nabla^* C(t)C(t)^*\nabla V_t(x) + x^*A(t)^*\nabla V_t(x) - \frac{1}{4}\|Q(t)^{-1/2}B(t)^*\nabla V_t(x)\|^2 + x^*P(t)x.$$

Let us try a value function of the form $V_t(x) = x^*F(t)x + g(t)$, where $F(t)$ is a time-dependent $n \times n$ symmetric matrix and $g(t)$ is a scalar function. Straightforward computation gives

$$\frac{d}{dt}F(t) + A(t)^*F(t) + F(t)A(t) - F(t)B(t)Q(t)^{-1}B(t)^*F(t) + P(t) = 0,$$
$$\frac{d}{dt}g(t) + \mathrm{Tr}[C(t)^*F(t)C(t)] = 0,$$

with the terminal conditions $F(T) = R$ and $g(T) = 0$, and the associated candidate policy

$$\alpha^*(t,x) = -Q(t)^{-1}B(t)^*F(t)x.$$

By well known properties of the matrix Riccati equation, see [Won68a, theorem 2.1], and our assumptions on the various matrices that appear in the problem, the equation for $F(t)$ has a unique C^1 solution on $[0,T]$. Hence the coefficients of the controlled equation for $X_t^{u^*}$, with $u_t^* = \alpha^*(t, X_t^{u^*})$, are uniformly Lipschitz continuous, and thus by proposition 6.2.1 and theorem 5.1.3 all the requirements for verification are satisfied. Thus we are done. □

We can also investigate the linear regulator on the infinite time horizon. Let us investigate the time-average cost (the discounted problem can also be solved, but this is less common in applications). To this end, we consider the time-homogeneous case,

$$dX_t^u = AX_t^u\,dt + Bu_t\,dt + C\,dW_t,$$

with the associated time-average cost functional

$$J_\infty[u] = \limsup_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\int_0^T \{(X_t^u)^* P X_t^u + (u_t)^* Q u_t\}\,dt\right].$$

Let us make the following additional assumptions.

1. $\mathbb{E}(\|X_0\|^2) < \infty$;
2. P and Q are symmetric matrices (they can always be symmetrized);
3. P is positive semidefinite and Q is positive definite;
4. (A,B) and $(A^*, \sqrt{P})$ are stabilizable.

Recall that a pair of matrices (A,B) is called *stabilizable* if there exists a matrix K (with the appropriate dimensions) such that all the eigenvalues of the matrix $A-BK$ have negative real parts. Conditions for this to be the case can be found in any good book on linear systems theory, see, e.g., [KS72].

Our goal is to find a control strategy that minimizes $J_\infty[u]$.

Theorem 6.5.2 (Linear regulator, time-average cost). *Let F be a positive semidefinite solution of the algebraic Riccati equation $A^*F + FA - FBQ^{-1}B^*F + P = 0$. Then $u_t^* = -Q^{-1}B^*FX_t^u$ is an optimal control for the cost $J_\infty[u]$.*

Proof. The Bellman equation in the time-average setting becomes

$$0 = \min_{\alpha\in\mathbb{R}^k}\left\{(Ax+B\alpha)^*\nabla V(x) + \frac{1}{2}\nabla^* CC^*\nabla V(x) + x^*Px + \alpha^*Q\alpha - \eta\right\},$$

As Q is positive definite, the minimum is attained at

$$\alpha^*(x) = -\frac{1}{2}Q^{-1}B^*\nabla V(x),$$

so the Bellman equation can be written as

$$0 = \frac{1}{2}\nabla^* CC^*\nabla V(x) + x^*A^*\nabla V(x) - \frac{1}{4}\|Q^{-1/2}B^*\nabla V(x)\|^2 + x^*Px - \eta.$$

Let us try a function of the form $V(x) = x^* F x$, where F is an $n \times n$ symmetric matrix. Then

$$A^*F + FA - FBQ^{-1}B^*F + P = 0, \qquad \eta = \mathrm{Tr}[C^*FC],$$

and the associated candidate policy becomes $\alpha^*(x) = -Q^{-1}B^*Fx$. We now invoke the properties of the algebraic Riccati equation, see [Won68a, theorem 4.1]. By the stabilizability assumption, there is at least one positive semidefinite solution F, such that $A - BQ^{-1}B^*F$ is a stable matrix. Using the latter and the controlled equation for $X_t^{u^*}$ with $u_t^* = \alpha^*(X_t^{u^*})$, you can verify by explicit computation that $\mathbb{E}(V(X_t^{u^*}))$ is bounded in time. Thus the asymptotic condition for verification is satisfied, while the martingale condition is clearly satisfied by theorem 5.1.3. Thus we find that u_t^* is indeed an optimal control. □

6.6 Markov chain approximation

Just like most stochastic differential equations do not admit analytic solution, most stochastic control problems can not be solved analytically either. It is thus of interest to develop numerical methods that can be used to solve such problems; otherwise we are essentially restricted to the linear regulator (which is, however, widely used in applications) and a small selection of other special cases. One could argue that the numerical solution of stochastic control problems essentially boils down to the numerical solution of a highly nonlinear PDE, the Bellman equation. This is indeed one way of looking at the problem, but there is a much more probabilistic approach that one can take as well. The goal of this section is to outline the latter method through a couple of simple examples. A proof of convergence will unfortunately be beyond our scope, but ample discussion of this can be found in the literature.

Remark 6.6.1. Stochastic optimal control problems suffer from the *curse of dimensionality*, as do most problems that require the numerical computation of a function on a high-dimensional state space. In low dimensions (e.g., one through three are often doable) one can numerically evaluate the function on a suitably selected grid (for finite-difference type schemes) or mesh (as in finite element methods), but the complexity of such a discretization will grow exponentially with the dimension of the state space. As such these methods quickly become intractable in higher dimensions, unless some additional structure can be taken into account to simplify the problem.

Let us reexamine the indefinite time tracking problem of example 6.3.2. For sake of demonstration, we will develop a numerical method to solve this problem; as we have already solved the problem analytically, we will be able to check the precision of the numerical method. Recall that the Bellman equation for this problem is given by

$$0 = \min_{\alpha \in \mathbb{U}} \left\{ \frac{\sigma^2}{2} \frac{\partial^2 V(x)}{\partial x^2} + \beta\alpha \frac{\partial V(x)}{\partial x} + p\alpha^2 - q \right\},$$

with the boundary conditions $V(r) = V(-r) = 0$. To approximate this equation, let us discretize the interval $[-r, r]$ into a grid $S_\delta = \{kr/N : k = -N, \ldots, N\}$ for some $N \in \mathbb{N}$. For notational simplicity, we denote by $\delta = r/N$ the spacing between the grid points. We now introduce the following finite-difference approximations:

$$\frac{\partial^2 V(x)}{\partial x^2} \approx \frac{V(x+\delta) - 2V(x) + V(x-\delta)}{\delta^2}, \qquad \frac{\partial V(x)}{\partial x} \approx \frac{V(x+\delta) - V(x-\delta)}{2\delta}$$

for $x \in S'_\delta = S_\delta \backslash \{-r, r\}$ (the interior of S_δ). This particular choice for the discretization of the differential operators is not arbitrary: we will shortly see that the careful choice of discretization results in a particularly sensible approximation.

Let us call the approximate value function $V_\delta(x)$. Then

$$0 = \min_{\alpha \in \mathbb{U}_\delta} \left\{ \frac{\sigma^2}{2} \frac{V_\delta(x+\delta) - 2V_\delta(x) + V_\delta(x-\delta)}{\delta^2} + \beta\alpha \frac{V_\delta(x+\delta) - V_\delta(x-\delta)}{2\delta} + p\alpha^2 - q \right\}, \qquad x \in S'_\delta,$$

which becomes after a little rearranging

$$V_\delta(x) = \min_{\alpha \in \mathbb{U}_\delta} \left\{ \frac{1}{2}(V_\delta(x+\delta) + V_\delta(x-\delta)) + \frac{\beta\alpha\delta}{2\sigma^2}(V_\delta(x+\delta) - V_\delta(x-\delta)) + \frac{p\alpha^2\delta^2}{\sigma^2} - \frac{q\delta^2}{\sigma^2} \right\}, \qquad x \in S'_\delta,$$

where for $x \notin S'_\delta$ we obviously choose the boundary conditions $V_\delta(r) = V_\delta(-r) = 0$. Let us now define the $(2N-1) \times (2N-1)$-dimensional matrix P^α with entries

$$P^\alpha_{i,i+1} = \frac{1}{2} + \frac{\beta\alpha\delta}{2\sigma^2}, \qquad P^\alpha_{i,i-1} = \frac{1}{2} - \frac{\beta\alpha\delta}{2\sigma^2}, \qquad P^\alpha_{i,j} = 0 \text{ for } j \neq i+1, i-1.$$

Provided that we choose our approximate control set $\mathbb{U}_\delta \subset [-\sigma^2/\beta\delta, \sigma^2/\beta\delta]$, we see that the entries of P^α are nonnegative and $\sum_j P^\alpha_{i,j} = 1$ for $j \neq 1, 2N-1$. Evidently, *P^α is the transition probability matrix for a discrete time Markov chain x^α_n with values in S_δ and with absorbing boundaries*. But there is more: as we show next, our approximation to the Bellman equation is itself the dynamic programming equation for an optimal control problem for the Markov chain x^α_n! Hence our finite-difference approximation is much more than an approximation to a PDE: *it approximates our entire control problem by a new (discretized) optimal control problem.*

Proposition 6.6.2. *Denote by x^u_n the controlled Markov chain on S_δ with*

$$\mathbb{P}(x^u_n = (k \pm 1)\delta | x^u_{n-1} = k\delta) = \frac{1}{2} \pm \frac{\beta\delta}{2\sigma^2}\,\alpha(n, k\delta), \quad k = -N+1, \ldots, N-1,$$
$$\mathbb{P}(x^u_n = \pm r | x^u_{n-1} = \pm r) = 1, \qquad \mathbb{P}(x_0 \in S'_\delta) = 1,$$

where the feedback control strategy u is assumed to be of the Markov type $u_n = \alpha(n, x^u_{n-1})$. Denote by $\sigma^u = \min\{n : x^u_n = \pm r\}$, and introduce the cost functional

$$K[u] = \mathbb{E}\left[\sum_{n=1}^{\sigma^u} (p\,(u_n)^2 - q)\,\frac{\delta^2}{\sigma^2} \right].$$

Denote by $V_\delta(x)$ the solution to the equation above, by $\alpha^*(x)$ the associated minimum, and $u^*_n = \alpha^*(x^{u^*}_{n-1})$. If $\mathbb{E}(\sigma^{u^*}) < \infty$, then $K[u^*] \leq K[u]$ for any Markov control u with values in $\mathbb{U}_\delta$ such that $\mathbb{E}(\sigma^u) < \infty$. Moreover, we can write $K[u^*] = \mathbb{E}(V_\delta(x_0))$.

The proof should look vaguely familiar!

Proof. For $x \in S'_\delta$, note that we can write

$$\mathbb{E}(V_\delta(x^u_n)|x^u_{n-1} = x) = \frac{1}{2}(V_\delta(x+\delta) + V_\delta(x-\delta)) + \frac{\beta\delta}{2\sigma^2}\,\alpha(n,x)\,(V_\delta(x+\delta) - V_\delta(x-\delta)).$$

Hence we obtain, using the equation for $V_\delta(x)$ on $x \in S'_\delta$,

$$\mathbb{E}(V_\delta(x^u_{n-1}) - V_\delta(x^u_n)|x^u_{n-1} = x) \le (p(\alpha(n,x))^2 - q)\,\frac{\delta^2}{\sigma^2}.$$

Multiplying both sides by $I_{x\in S'_\delta}$, setting $x = x^u_{n-1}$ and taking the expectation, we find that

$$\mathbb{E}\left((V_\delta(x^u_{n-1}) - V_\delta(x^u_n))\,I_{x^u_{n-1}\in S'_\delta}\right) \le \mathbb{E}\left[I_{x^u_{n-1}\in S'_\delta}\,(p(u_n)^2 - q)\,\frac{\delta^2}{\sigma^2}\right].$$

Summing over n up to some $T \in \mathbb{N}$, we find that

$$\mathbb{E}\left(\sum_{n=1}^{T}(V_\delta(x^u_{n-1}) - V_\delta(x^u_n))\,I_{x^u_{n-1}\in S'_\delta}\right) \le \mathbb{E}\left[\sum_{n=1}^{T} I_{x^u_{n-1}\in S'_\delta}\,(p(u_n)^2 - q)\,\frac{\delta^2}{\sigma^2}\right],$$

or, rewriting this expression in a familiar form,

$$\mathbb{E}(V_\delta(x_0)) \le \mathbb{E}\left[V_\delta(x^u_{T\wedge\sigma^u}) + \sum_{n=1}^{T\wedge\sigma^u}(p(u_n)^2 - q)\,\frac{\delta^2}{\sigma^2}\right].$$

As $\mathbb{U}_\delta$ is bounded and as $\mathbb{E}(\sigma^u) < \infty$, we can now let $T \to \infty$ by dominated convergence to obtain $\mathbb{E}(V_\delta(x_0)) \le K[u]$. But repeating the same arguments with u^*, we find that $\mathbb{E}(V_\delta(x_0)) = K[u^*]$. Thus u^* is indeed an optimal strategy, and the proof is complete. □

Now that we have discretized our control problem, how do we solve the discrete problem? There are various ways of doing this, many of which are detailed in the books [Kus71, KD01]. One of the simplest is the *Jacobi method*, which works as follows. Start with an arbitrary choice for $V^0_\delta(x)$. Now define, for any $n \in \mathbb{N}$,

$$V^n_\delta(x) = \min_{\alpha\in\mathbb{U}_\delta}\left\{\frac{1}{2}(V^{n-1}_\delta(x+\delta) + V^{n-1}_\delta(x-\delta))\right.$$
$$\left. + \frac{\beta\alpha\delta}{2\sigma^2}(V^{n-1}_\delta(x+\delta) - V^{n-1}_\delta(x-\delta)) + \frac{p\alpha^2\delta^2}{\sigma^2} - \frac{q\delta^2}{\sigma^2}\right\}, \qquad x \in S'_\delta,$$

where we impose the boundary conditions $V^{n-1}_\delta(\pm r) = 0$ for every n. The minimum in this iteration is easily seen to be attained at (setting $\mathbb{U}_\delta = [-\sigma^2/\beta\delta, \sigma^2/\beta\delta]$)

$$\alpha^*_{\delta,n}(x) = \left(-\frac{\beta}{4p\delta}(V^{n-1}_\delta(x+\delta) - V^{n-1}_\delta(x-\delta))\right) \vee \left(-\frac{\sigma^2}{\beta\delta}\right) \wedge \frac{\sigma^2}{\beta\delta}.$$

It is not difficult to prove that the iteration for $V^n_\delta(x)$ will converge to some function $V_\delta(x)$ as $n \to \infty$, see [Kus71, theorem 4.4], and this limit is indeed the value function for our approximate optimal control problem, while the limit of $\alpha^*_{\delta,n}(x)$ as $n \to \infty$ is the optimal control for our approximate optimal control problem. Other methods

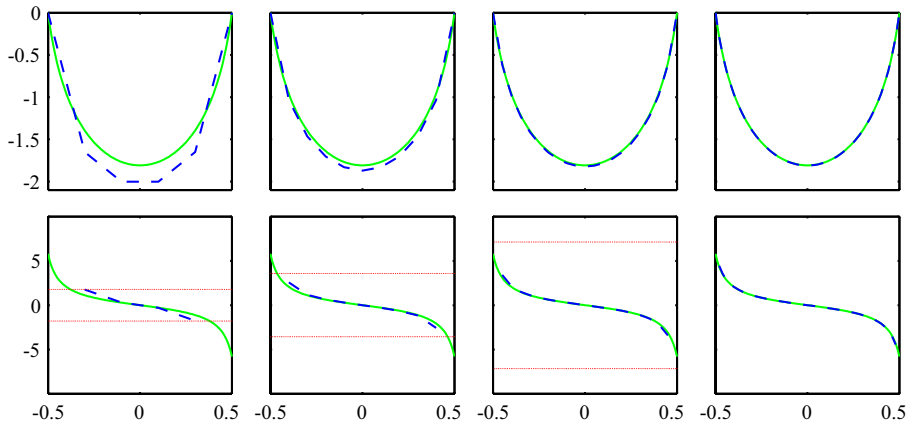

Figure 6.1. Numerical solution of example 6.3.2 with $r = .5$, $\beta = .7$, $p = q = 1$, and $\sigma = .5$. From left to right, the interval was discretized into 6, 11, 21 and 51 points. The top plots show $V_\delta(x)$ (dashed line) and the analytic solution for $V(x)$ (solid line). The bottom plots show the discrete optimal strategy $\alpha^*_\delta(x)$ (dashed line) and the analytic solution for $\alpha^*(x)$ (solid line). The dotted horizontal lines are the upper and lower bounds on the discrete control set $\mathbb{U}_\delta$.

often converge faster (e.g., the Gauss-Seidel method [Kus71, theorem 4.6]) and are not much more difficult to implement, but the Jacobi method will do for our purposes.

The result of implementing this procedure on a computer is shown in figure 6.1, together then the analytical solution obtained in example 6.3.2, for a particular choice of parameters. For all but the coarsest discretization, both the discretized value function and the optimal control are quite close to their analytic solutions; in fact, it appears that not too fine a grid already gives excellent performance!

Remark 6.6.3. We will not give a proof of convergence here, but you can imagine why some form of Markov chain approximation would be a good thing to do. The convergence proof for this procedure does not rely at all on showing that the solution $V_\delta(x)$ converges to the solution $V(x)$ of the continuous Bellman equation. Instead, one proceeds by showing that the Markov chain x^u_n converges as $\delta \to 0$, in a suitable sense, to the controlled diffusion X^u_t. One can then show that the optimal control policy for x^u_n also converges, in a suitable sense, to an optimal control policy for the diffusion X^u_t, without invoking directly the continuous time verification theorems. The fact that all the objects in these approximations are probabilistic—and that every discretized problem is itself an optimal control problem—is thus a key (and quite nontrivial) idea. For detailed references on this topic, see section 6.7.

Remark 6.6.4. The finite-difference method is only a tool to obtain a suitable Markov chain approximation for the original problem. The fact that this approximation has its origins in a finite-difference method is not used in the convergence proofs; in fact, any Markov chain that satisfies a set of "local consistency" conditions suffices, though some approximations will converge faster (as $\delta \to 0$) than others. Even the finite-

difference scheme is not unique: there are many such schemes that give rise to Markov chain approximations (though there are also many which do not!). In particular, one can obtain an approximation without the constraint on $\mathbb{U}_\delta$ by using one-sided differences for the derivatives, provided their sign is chosen correctly; on the other hand, the central difference approximation that we have used is known to converge faster in most cases (see [KD01, chapter 5] for details).

To demonstrate the method further, let us discuss another very simple example.

Example 6.6.5 (Inverted pendulum). Consider a simple pendulum in one dimension, which experiences random forcing and is heavily damped. We use the model

$$d\theta_t^u = c_1 \sin(\theta_t^u)\, dt - c_2 \cos(\theta_t^u)\, u_t\, dt + \sigma\, dW_t,$$

where θ_t is the angle relative to the up position ($\theta = 0$), $c_1, c_2, \sigma > 0$ are constants, and we have allowed for a control input u_t of the "pendulum on a cart" type (the control is ineffective when the pendulum is horizontal). Starting in the down position ($\theta = \pi$), we would like to flip the pendulum to the up position as quickly as possible—with an angular precision $\varepsilon > 0$, say—while minimizing the total control power necessary to achieve this task. We thus introduce the stopping time and cost

$$\tau^u = \inf\{t : \theta_t^u \le \varepsilon \text{ or } \theta_t^u \ge 2\pi - \varepsilon\}, \qquad J[u] = \mathbb{E}\left[\int_0^{\tau^u} \{p\,(u_s)^2 + q\}\, ds\right],$$

where $p, q > 0$ are constants that determine the tradeoff between minimizing the inversion time and minimizing the necessary power. The Bellman equation becomes

$$0 = \min_{\alpha\in\mathbb{U}} \left\{ \frac{\sigma^2}{2}\frac{\partial^2 V(x)}{\partial x^2} + (c_1 \sin(x) - c_2 \cos(x)\,\alpha)\,\frac{\partial V(x)}{\partial x} + p\alpha^2 + q \right\}$$

for $x \in \,]\varepsilon, 2\pi - \varepsilon[$, with the boundary conditions $V(\varepsilon) = V(2\pi - \varepsilon) = 0$.

We proceed to approximate the Bellman equation using the same finite-difference approximation used in the previous example. This gives, after some manipulation,

$$\begin{aligned} V_\delta(x) = \min_{\alpha\in\mathbb{U}_\delta} \Bigg\{ &\frac{1}{2}(V_\delta(x+\delta) + V_\delta(x-\delta)) + \frac{p\alpha^2\delta^2}{\sigma^2} + \frac{q\delta^2}{\sigma^2} \\ &+ \frac{\delta}{2\sigma^2}(c_1 \sin(x) - c_2 \cos(x)\,\alpha)(V_\delta(x+\delta) - V_\delta(x-\delta)) \Bigg\}, \qquad x \in S'_\delta, \end{aligned}$$

where we have set $\delta = (\pi - \varepsilon)/N$, $S_\delta = \{\pi + k(\pi-\varepsilon)/N : k = -N, \dots, N\}$, $S'_\delta = S_\delta \backslash \{\varepsilon, 2\pi - \varepsilon\}$, and we impose the boundary conditions $V_\delta(\varepsilon) = V_\delta(2\pi - \varepsilon) = 0$. We now need to choose the control interval $\mathbb{U}_\delta$ so that the coefficients in the approximate Bellman equation are transition probabilities, i.e., we need to make sure that

$$\frac{\delta}{2\sigma^2}\,|c_1 \sin(x) - c_2 \cos(x)\,\alpha| \le \frac{1}{2} \qquad \forall\, \alpha \in \mathbb{U}_\delta.$$

For example, we can set $\mathbb{U}_\delta = [-G, G]$ with $G = \sigma^2/c_2\delta - c_1/c_2$, provided that we require δ to be sufficiently small that the constant G is positive.

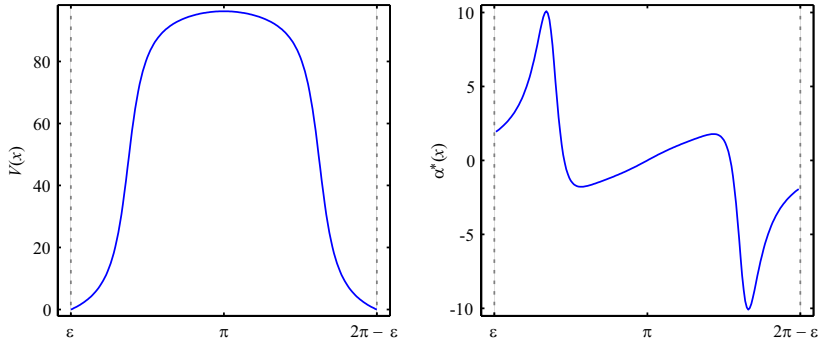

Figure 6.2. Numerical solution of example 6.6.5 with $\varepsilon = .25$, $c_1 = c_2 = .5$, $p = q = 1$, and $\sigma = .5$. The interval was discretized into 201 points ($N = 100$). The left plot shows the value function $V(x)$; the right plot shows the optimal control $\alpha^*(x)$.

Next, we define the Jacobi iterations, starting from any $V_\delta^0(x)$, by

$$V_\delta^n(x) = \min_{\alpha \in \mathbb{U}_\delta} \left\{ \frac{1}{2}(V_\delta^{n-1}(x+\delta) + V_\delta^{n-1}(x-\delta)) + \frac{p\alpha^2\delta^2}{\sigma^2} + \frac{q\delta^2}{\sigma^2} \right.$$
$$\left. + \frac{\delta}{2\sigma^2}(c_1 \sin(x) - c_2 \cos(x)\,\alpha)(V_\delta^{n-1}(x+\delta) - V_\delta^{n-1}(x-\delta)) \right\},$$

and note that the minimum is in fact attained at

$$\alpha^*_{\delta,n}(x) = \left(\frac{c_2}{4p\delta} \cos(x)\,(V_\delta^{n-1}(x+\delta) - V_\delta^{n-1}(x-\delta)) \right) \vee (-G) \wedge G.$$

Little remains but to implement the method, the result of which is shown in figure 6.2.

Remark 6.6.6. We have only discussed approximation of the indefinite time control problems in one dimension. The method extends readily to multiple dimensions, provided (as always) that sufficient care is taken to choose appropriate finite differences, and that sufficient computational power is available. This type of method is also extremely flexible in that it extends to a wide variety of control problems, and is certainly not restricted to the indefinite time problem. However, the latter has the nice property that it is naturally restricted to a bounded domain. For other costs this need not be the case, so that one has to take care to truncate the state space appropriately. Of course, any grid-based numerical method will suffer from the same problem.

6.7 Further reading

There are several good textbooks on stochastic optimal control in continuous time. For detailed treatments of the subject, check out the books by Fleming and Rishel

[FR75], Fleming and Soner [FS06], Yong and Zhou [YZ99] or Krylov [Kry80]. A recent review article with many further references is Borkar [Bor05]. A nice recent book with a strong emphasis on verification theorems is Øksendal and Sulem [ØS05]. Finally, Hanson [Han07] gives a non-mathematical introduction to the subject with an emphasis on applications and computational methods.

Readers familiar with optimal control in the deterministic setting would likely be quick to point out that dynamic programming is not the only way to go; in fact, methods based on Pontryagin's maximum principle are often preferable in the deterministic setting. Such methods also exist in the stochastic case; see Yong and Zhou [YZ99] for an extensive discussion and for further references. To date, the dynamic programming approach has been more successful in the stochastic case than the maximum principle approach, if only for technical reasons. The maximum principle requires the solution of an SDE with a terminal condition rather than an initial condition, but whose solution is nonetheless adapted—a feat that our SDE theory certainly cannot accomplish! On the other hand, the dynamic programming approach requires little more than the basic tools of the trade, at least in the simplest setting (as we have seen).

The martingale dynamic programming principle gives a rather attractive probabilistic spin to the dynamic programming method; it is also useful in cases where there is insufficient regularity (the value function is not "nice enough") for the usual approach to work. A lucid discussion of the martingale approach can be found in the book by Elliott [Ell82]; an overview is given by Davis in [Dav79].

Lemma 6.3.3, which guarantees that the exit time from a bounded set has finite expectation (under a nondegeneracy condition), is taken from [Has80, section III.7].

Robin [Rob83] gives a nice overview of stochastic optimal control problems with time-average cost. The book by Davis [Dav77] gives an excellent introduction to linear stochastic control theory (i.e., the linear regulator and its relatives).

Markov chain approximations in stochastic control are developed extensively in the books by Kushner [Kus77] and by Kushner and Dupuis [KD01]. A nice overview can be found in Kushner [Kus90]. In this context, it is important to understand the optimal control of discrete time, discrete state space Markov chains; this is treated in detail in Kushner [Kus71] and in Kumar and Varaiya [KV86]. Kushner and Dupuis [KD01] and Kushner [Kus71] detail various algorithms for the solution of discrete stochastic optimal control problems, including the simple by effective Jacobi and Gauss-Seidel methods. The convergence proofs for the Markov chain approximation itself rely heavily on the theory of weak convergence, see the classic book by Billingsley [Bil99], Ethier and Kurtz [EK86], and yet another book by Kushner [Kus84].

A different approach to numerical methods for stochastic optimal control problems in continuous time is direct approximation of the Bellman PDE. Once a suitable numerical method has been obtained, one can then attempt to prove that its solution converges in some sense to a solution (in the viscosity sense) of the continuous Bellman equation. See, for example, the last chapter of Fleming and Soner [FS06].

One of the most intriguing aspects of optimal stochastic control theory is that it can sometimes be applied to obtain results in other, seemingly unrelated, areas of mathematics. Some selected applications can be found in Borell [Bor00] (geometric analysis), Sheu [She91] (heat kernel estimates), Dupuis and Ellis [DE97] and Boué and Dupuis [BD98] (large deviations), Fleming and Soner [FS06] (singular perturba-

tion methods), and in Fleming and Mitter [FM83] and Mitter and Newton [MN03] (nonlinear filtering). Dupuis and Oliensis [DO94] discuss an interesting application to three-dimensional surface reconstruction from a two-dimensional image.

To date, the most important areas of application for optimal stochastic control are mathematical finance and engineering. An excellent reference for financial applications is the well-known book by Karatzas and Shreve [KS98]. In engineering, the most important part of the theory remains (due to the fact that it is tractable) stochastic control of linear systems. However, this theory goes far beyond the linear regulator; for one example of this, see the recent article by Petersen [Pet06].

CHAPTER 7

Filtering Theory

Filtering theory is concerned with the following problem. Suppose we have some *signal process*—a stochastic process X_t—which we cannot observe directly. Instead, we are given an *observation process* Y_t which is correlated with X_t; we will restrict ourselves to the important special case of "signal plus white noise" type observations $dY_t = h(X_t)\,dt + \sigma\,dW_t$, where W_t is a Wiener process. Given that by time t we can only observe $\{Y_s : s \leq t\}$, it becomes necessary to *estimate* X_t from the observations $Y_{s\leq t}$. For any function f, we have already seen that the *best* estimate, in the mean square sense, of $f(X_t)$ given $Y_{s\leq t}$, is given by the conditional expectation $\pi_t(f) \equiv \mathbb{E}(f(X_t)|\mathcal{F}_t^Y)$, where $\mathcal{F}_t^Y = \sigma\{Y_s : s \leq t\}$ (see proposition 2.3.3).

The goal of the filtering problem is to find an explicit expression for $\pi_t(f)$ in terms of $Y_{s\leq t}$; in particular, we will seek to express $\pi_t(f)$ as the solution of a *stochastic differential equation* driven by Y_t. This is interesting in itself: it leads to algorithms that allow us to optimally estimate a signal in white noise, which is important in many applications. In addition, we will see that filtering also forms an integral part of stochastic optimal control in the case where the feedback signal in only allowed to depend on noisy observations (which is the case in many applications), rather than assuming that we can precisely observe the state of the system (which we have done throughout the previous chapter). Before we can tackle any of these problems, however, we need to take a closer look at some of the properties of the conditional expectation.

7.1 The Bayes formula

In your undergraduate probability course, you likely encountered two types of conditioning. The first type is the calculation of conditional expectations for finite-valued random variables; this idea was developed in section 2.1, and we have seen that it is a special case of the general definition of the conditional expectation. The second ap-

proach is for continuous random variables with probability densities; as we have not yet discussed conditional expectations in this setting, let us take a moment to show how this idea relates to the general definition of the conditional expectation.

Example 7.1.1 (Conditioning with densities). Consider two random variables X, Y on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, such that X and Y both take values in the interval $[0,1]$. We will assume that X and Y possess as *joint density* $p(x,y)$*;* by this we mean that for any bounded measurable function $f : [0,1] \times [0,1] \to \mathbb{R}$, we can write

$$\mathbb{E}(f(X,Y)) = \int_0^1 \int_0^1 f(x,y)\, p(x,y)\, dx\, dy.$$

In your undergraduate course, you likely learned that the "conditional expectation" of $f(X,Y)$, given $Y = y$ ($y \in [0,1]$), is given by

$$\mathbb{E}(f(X,Y)|Y=y) = \frac{\int_0^1 f(x,y)\, p(x,y)\, dx}{\int_0^1 p(x,y)\, dx}.$$

This is often justified by analogy with the discrete case: if X, Y take discrete values $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$, rather than continuous values $x, y \in [0,1]$, then (why?)

$$\mathbb{E}(f(X,Y)|Y=y_j) = \frac{\sum_{i=1}^m f(x_i, y_j)\, p_{ij}}{\sum_{i=1}^m p_{ij}}, \qquad p_{ij} = \mathbb{P}(X = x_i \text{ and } Y = y_j).$$

On the other hand, it is not at all clear that the quantity $\mathbb{E}(f(X,Y)|Y=y)$ is even meaningful in the continuous case—after all, $\mathbb{P}(Y = y) = 0$ for any $y \in [0,1]$! (This is necessarily true as, by our assumption that $p(x,y)$ exists, the law of Y is absolutely continuous with respect to the uniform measure on $[0,1]$.)

To make mathematical sense of this construction, consider the more meaningful expression (which is similarly the natural analog of the discrete case)

$$\mathbb{E}(f(X,Y)|Y) = \frac{\int_0^1 f(x,Y)\, p(x,Y)\, dx}{\int_0^1 p(x,Y)\, dx} \equiv M_f(Y).$$

We claim that the random variable $M_f(Y)$, defined in this way, does indeed satisfy the Kolmogorov definition of the conditional expectation. To verify this, it suffices to show that $\mathbb{E}(\mathbb{E}(f(X,Y)|Y)\, u(Y)) = \mathbb{E}(f(X,Y)\, u(Y))$ for any bounded measurable u (after all, the indicator function I_A is of this form for any $A \in \sigma\{Y\}$). But

$$\begin{aligned}\mathbb{E}(M_f(Y)\, u(Y)) &= \int_0^1 \int_0^1 M_f(y)\, u(y)\, p(x,y)\, dx\, dy \\ &= \int_0^1 \frac{\int_0^1 f(x,y)\, p(x,y)\, dx}{\int_0^1 p(x,y)\, dx}\, u(y) \left[\int_0^1 p(x,y)\, dx\right] dy = \mathbb{E}(f(X,Y)\, u(Y)),\end{aligned}$$

which is precisely what we set out to show.

A trivial but particularly interesting case of this construction occurs for the density $p(x,y) = p(x)q(y)$, i.e., when X and Y are independent, X has density $p(x)$ and Y has density $q(y)$. In this case, we find that the conditional expectation is given by

$$\mathbb{E}(f(X,Y)|Y) = \int_0^1 f(x,Y)\, p(x)\, dx.$$

Evidently, when two random variables are independent, conditioning on one of the random variables simply corresponds to averaging over the other. You should convince yourself that intuitively, this makes perfect sense! In fact, we have already used this idea in disguise: have another look at the proof of lemma 3.1.9.

The conclusion of the previous example provides a good excuse for introductory courses not to introduce measure theory as the cornerstone of probability. However, the fundamental idea that underlies this example becomes much more powerful (and conceptually clear!) when interpreted in a measure-theoretic framework. Let us thus repeat the previous example, but from a measure-theoretic point of view.

Example 7.1.2 (Conditioning with densities II). Consider the space $\Omega = [0,1] \times [0,1]$, endowed with its Borel σ-algebra $\mathcal{F} = \mathcal{B}([0,1]) \times \mathcal{B}([0,1])$ and some probability measure $\mathbb{P}$. Denote by $Y : \Omega \to [0,1]$ the canonical random variable $Y(x,y) = y$, and let Z be any integrable random variable on Ω. Beside $\mathbb{P}$, we introduce also the product measure $\mathbb{Q} = \mu_0 \times \mu_0$, where μ_0 is the uniform measure on $[0,1]$.

Now suppose that the measure $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$. Then

$$\mathbb{E}_{\mathbb{P}}(Z) = \mathbb{E}_{\mathbb{Q}}\left(Z\,\frac{d\mathbb{P}}{d\mathbb{Q}}\right) = \int_0^1 \int_0^1 Z(x,y)\,\frac{d\mathbb{P}}{d\mathbb{Q}}(x,y)\,dx\,dy,$$

where we have expressed the uniform measure in the usual calculus notation. Clearly $d\mathbb{P}/d\mathbb{Q}$ is the density $p(x,y)$ of the previous example, and (by the Radon-Nikodym theorem) the existence of $p(x,y)$ is precisely the requirement that $\mathbb{P} \ll \mathbb{Q}$.

We now have two probability measures $\mathbb{P}$ and $\mathbb{Q}$. Ultimately, we are interested in computing the conditional expectation $\mathbb{E}_{\mathbb{P}}(Z|Y)$. It is not immediately obvious how to do this! On the other hand, under the measure $\mathbb{Q}$, the computation of the conditional expectation $\mathbb{E}_{\mathbb{Q}}(Z|Y)$ is particularly simple. Let us consider this problem first. We claim that for any integrable random variable Z (i.e., $\mathbb{E}_{\mathbb{Q}}(|Z|) < \infty$), we can write

$$\mathbb{E}_{\mathbb{Q}}(Z|Y)(x,y) = \int_{[0,1]} Z(x,y)\,\mu_0(dx) = \int_0^1 Z(x,y)\,dx.$$

To be precise, we should first verify that this random variable is in fact measurable—this is indeed the case by Fubini's theorem. Let us now check Kolmogorov's definition of the conditional expectation. First, note that $\sigma\{Y\} = \{Y^{-1}(A) : A \in \mathcal{B}([0,1])\} = \{[0,1] \times A : A \in \mathcal{B}([0,1])\}$. Hence for any $S \in \sigma\{Y\}$, the indicator function

$I_S(x,y) = I_S(y)$ is only a function of y. Therefore, we find that for any $S \in \sigma\{Y\}$,

$$\mathbb{E}_{\mathbb{Q}}(I_S\,\mathbb{E}_{\mathbb{Q}}(Z|Y)) = \int_\Omega \left\{ I_S(y) \int_{[0,1]} Z(x,y)\,\mu_0(dx) \right\} \mathbb{Q}(dx,dy) = \int_{[0,1]} \left\{ I_S(y) \int_{[0,1]} Z(x,y)\,\mu_0(dx) \right\} \mu_0(dy) = \mathbb{E}_{\mathbb{Q}}(I_S\,Z),$$

where we have used Fubini's theorem to write the repeated integral as a product integral. But this is precisely the Kolmogorov definition—so the claim is established.

Apparently the computation of $\mathbb{E}_{\mathbb{Q}}(Z|Y)$ is more or less trivial. If only we were interested in the measure $\mathbb{Q}$! Unfortunately, in real life we are interested in the measure $\mathbb{P}$, which could be much more complicated than $\mathbb{Q}$ (and is most likely not a product measure). Now, however, we have a cunning idea. As $\mathbb{P} \ll \mathbb{Q}$, we know that one can express expectations under $\mathbb{P}$ as expectations under $\mathbb{Q}$ by inserting the Radon-Nikodym derivative: $\mathbb{E}_{\mathbb{P}}(Z) = \mathbb{E}_{\mathbb{Q}}(Z\,d\mathbb{P}/d\mathbb{Q})$. Perhaps we can do the same with *conditional* expectations? In other words, we can try to express conditional expectations under $\mathbb{P}$ in terms of conditional expectations under $\mathbb{Q}$, which, one would think, should come down to dropping in Radon-Nikodym derivatives in the appropriate places. If we can make this work, then we can enjoy all the benefits of $\mathbb{Q}$: in particular, the simple formula for $\mathbb{E}_{\mathbb{Q}}(Z|Y)$ would apply, which is precisely the idea.

The question is thus: how do conditional expectations transform under a change of measure? Let us briefly interrupt our example develop the relevant result.

Lemma 7.1.3 (Bayes formula). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and $\mathbb{P} \ll \mathbb{Q}$ for some probability measure $\mathbb{Q}$. Then for any σ-algebra $\mathcal{G} \subset \mathcal{F}$ and for any integrable random variable X (i.e., we require $\mathbb{E}_{\mathbb{P}}(|X|) < \infty$), the Bayes formula holds:*

$$\mathbb{E}_{\mathbb{P}}(X|\mathcal{G}) = \frac{\mathbb{E}_{\mathbb{Q}}(X\,d\mathbb{P}/d\mathbb{Q}\,|\,\mathcal{G})}{\mathbb{E}_{\mathbb{Q}}(d\mathbb{P}/d\mathbb{Q}\,|\,\mathcal{G})} \qquad \mathbb{P}\text{-a.s.}$$

This expression is well-defined, as $\mathbb{P}(\mathbb{E}_{\mathbb{Q}}(d\mathbb{P}/d\mathbb{Q}\,|\,\mathcal{G}) = 0) = 0$.

Remark 7.1.4. Recall that the conditional expectation $\mathbb{E}_{\mathbb{P}}(X|\mathcal{G})$ is only defined up to $\mathbb{P}$-a.s. equivalence—hence we can never ask for a stronger result than one which holds $\mathbb{P}$-a.s. This is a more fundamental issue than a mere technicality. Note that $\mathbb{P} \ll \mathbb{Q}$ only states that $\mathbb{Q}(A) = 0$ implies $\mathbb{P}(A) = 0$, not necessarily the other way around. There could thus be two versions of $\mathbb{E}_{\mathbb{P}}(X|\mathcal{G})$ which are distinct with positive probability under the measure $\mathbb{Q}$! Evidently it is important to specify, when we deal with conditional expectations with respect to different measures, under which measure we are "almost sure" (a.s.) of our statements. Fortunately, in many applications (including the ones in this chapter) $\mathbb{Q} \ll \mathbb{P}$ as well, so that this is no longer an issue.

Proof. First, let us verify that the conditional expectations are defined, i.e., we need to check integrability. $\mathbb{E}_{\mathbb{P}}(|X|) < \infty$ by assumption, while $\mathbb{E}_{\mathbb{Q}}(|d\mathbb{P}/d\mathbb{Q}|) = 1$ (as $d\mathbb{P}/d\mathbb{Q}$ is nonnegative). Finally, $\mathbb{E}_{\mathbb{Q}}(|X\,d\mathbb{P}/d\mathbb{Q}|) = \mathbb{E}_{\mathbb{P}}(|X|) < \infty$, so all the quantities at least make sense.

The rest is essentially an exercise in using the elementary properties of the conditional expectation: you should verify that you understand all the steps! Let $S \in \mathcal{G}$ be arbitrary, and note that $\mathbb{E}_\mathbb{Q}(I_S\,\mathbb{E}_\mathbb{Q}(X\,d\mathbb{P}/d\mathbb{Q}\,|\,\mathcal{G})) = \mathbb{E}_\mathbb{Q}(I_S\,X\,d\mathbb{P}/d\mathbb{Q}) = \mathbb{E}_\mathbb{P}(I_S\,X) = \mathbb{E}_\mathbb{P}(I_S\,\mathbb{E}_\mathbb{P}(X|\mathcal{G})) = \mathbb{E}_\mathbb{Q}(I_S\,d\mathbb{P}/d\mathbb{Q}\,\mathbb{E}_\mathbb{P}(X|\mathcal{G})) = \mathbb{E}_\mathbb{Q}(I_S\,\mathbb{E}_\mathbb{Q}(d\mathbb{P}/d\mathbb{Q}\,|\,\mathcal{G})\,\mathbb{E}_\mathbb{P}(X|\mathcal{G}))$. But this holds for any $S \in \mathcal{G}$, so we find that $\mathbb{E}_\mathbb{Q}(X\,d\mathbb{P}/d\mathbb{Q}\,|\,\mathcal{G}) = \mathbb{E}_\mathbb{Q}(d\mathbb{P}/d\mathbb{Q}\,|\,\mathcal{G})\,\mathbb{E}_\mathbb{P}(X|\mathcal{G})$ $\mathbb{Q}$-a.s. (why?).

We would like to divide both sides by $\mathbb{E}_\mathbb{Q}(d\mathbb{P}/d\mathbb{Q}\,|\,\mathcal{G})$, so we must verify that this quantity is nonzero (it is clearly nonnegative). Define the set $S = \{\omega : \mathbb{E}_\mathbb{Q}(d\mathbb{P}/d\mathbb{Q}\,|\,\mathcal{G})(\omega) = 0\}$, and note that $S \in \mathcal{G}$ as the conditional expectation is $\mathcal{G}$-measurable by definition. Then $0 = \mathbb{E}_\mathbb{Q}(I_S\,\mathbb{E}_\mathbb{Q}(d\mathbb{P}/d\mathbb{Q}\,|\,\mathcal{G})) = \mathbb{E}_\mathbb{Q}(I_S\,d\mathbb{P}/d\mathbb{Q}) = \mathbb{P}(S)$. Hence we can go ahead with our division on the set S^c, which has unit probability under $\mathbb{P}$. The result follows directly. □

We can now complete our example. Using the Bayes formula, we find that $\mathbb{P}$-a.s.

$$\mathbb{E}_\mathbb{P}(Z|Y)(x,y) = \frac{\mathbb{E}_\mathbb{Q}(Z\,d\mathbb{P}/d\mathbb{Q}\,|\,Y)(x,y)}{\mathbb{E}_\mathbb{Q}(d\mathbb{P}/d\mathbb{Q}\,|\,Y)(x,y)} = \frac{\int_{[0,1]} Z(x,y)\,\frac{d\mathbb{P}}{d\mathbb{Q}}(x,y)\,\mu_0(dx)}{\int_{[0,1]} \frac{d\mathbb{P}}{d\mathbb{Q}}(x,y)\,\mu_0(dx)},$$

where we have substituted in our simple expression for $\mathbb{E}_\mathbb{Q}(Z|Y)$. But this is precisely the density expression for the conditional expectation! When viewed in this light, there is nothing particularly fundamental about the textbook example 7.1.1: it is simply a particular example of the behavior of the conditional expectation under an absolutely continuous change of measure. The new measure $\mu_0 \times \mu_0$, called the *reference measure*, is chosen for convenience; under the latter, the computation of the conditional expectations reduces to straightforward integration.

The generalization of example 7.1.1 to the measure-theoretic setting will pay off handsomely in the solution of the filtering problem. What is the benefit of abstraction? In general, we wish to calculate $\mathbb{E}_\mathbb{P}(X|\mathcal{G})$, where $\mathcal{G}$ need not be generated by a simple random variable—for example, in the filtering problem $\mathcal{G} = \sigma\{Y_s : s \le t\}$ is generated by an entire continuous path. On the space of continuous paths, the concept of a "density" in the sense of example 7.1.1 does not make sense; there is no such thing as the uniform measure (or even a Lebesgue measure) on the space of continuous paths! However, in our more abstract setup, we are free to choose any reference measure we wish; the important insight is that what really simplified example 7.1.1 was not the representation of the densities with respect to the uniform measure per se, but that under the uniform measure X and Y were *independent* (which allowed us to reduce conditioning under the reference measure to simple integration). In the general case, we can still seek a reference measure under which X and $\mathcal{G}$ are independent—we even already have the perfect tool for this purpose, the Girsanov theorem, in our toolbox waiting to be used! The abstract theory then allows us to proceed just like in example 7.1.1, even though we are no longer operating within its (restrictive) setting.

Example 7.1.5. Before developing a more general theory, let us demonstrate these ideas in the simplest filtering example: the estimation of a constant in white noise.

We work on the space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, \mathbb{P})$, on which are defined a (one-dimensional) $\mathcal{F}_t$-Wiener process W_t and an $\mathcal{F}_0$-measurable random variable X_0 (which is thus by definition independent of W_t). We consider a situation in which we cannot observe X_0 directly: we only have access to noisy observations of the form $y_t = X_0 + \kappa\,\xi_t$,

where ξ_t is white noise. As usual, we will work in practice with the integrated form of the observations to obtain a sensible mathematical model (see the Introduction and section 3.3 for discussion); i.e., we set $Y_t = X_0 t + \kappa W_t$. The goal of the filtering problem is to compute $\pi_t(f) \equiv \mathbb{E}_{\mathbb{P}}(f(X_0)|\mathcal{F}_t^Y)$, where $\mathcal{F}_t^Y = \sigma\{Y_s : s \le t\}$ is the *observation filtration*, for a sufficiently large class of functions f; then $\pi_t(f)$ is the optimal (least mean square) estimate of $f(X_0)$, given the observations up to time t.

To tackle this problem, we will essentially repeat example 7.1.1 in this setting. As we are conditioning on entire observation paths, we do not have a uniform measure available to us; nonetheless, we can find a reference measure $\mathbb{Q}$ under which X_0 and Y_t are *independent*, at least on finite time intervals! This is just Girsanov's theorem.

Lemma 7.1.6. *Suppose Λ_T^{-1} below satisfies $\mathbb{E}_{\mathbb{P}}(\Lambda_T^{-1}) = 1$, and define $\mathbb{Q}_T \ll \mathbb{P}$ by*

$$\frac{d\mathbb{Q}_T}{d\mathbb{P}} = \exp\left(-\kappa^{-1} X_0 W_T - \frac{1}{2}\kappa^{-2}(X_0)^2 T\right) \equiv \Lambda_T^{-1}.$$

Then under $\mathbb{Q}_T$ the random variable X_0 has the same law as under $\mathbb{P}$, and the process $\{\kappa^{-1} Y_t\}_{t\in[0,T]}$ is an $\mathcal{F}_t$-Wiener process independent of X_0. Moreover, $\mathbb{P} \ll \mathbb{Q}_T$ with

$$\frac{d\mathbb{P}}{d\mathbb{Q}_T} = \exp\left(\kappa^{-2} X_0 Y_T - \frac{1}{2}\kappa^{-2}(X_0)^2 T\right) = \Lambda_T.$$

Proof. By Girsanov's theorem (see also remark 4.5.4), $\{\kappa^{-1} Y_t\}_{t\in[0,T]}$ is an $\mathcal{F}_t$-Wiener process under $\mathbb{Q}_T$. But X_0 is $\mathcal{F}_0$-measurable, so $\{\kappa^{-1} Y_t\}_{t\in[0,T]}$ must be independent of X_0 under $\mathbb{Q}_T$. To show that X_0 has the same law under $\mathbb{Q}_T$, note that Λ_t^{-1} is a martingale under $\mathbb{P}$ (as it is a nonnegative local martingale, and thus a supermartingale, with constant expectation); hence $\mathbb{E}_{\mathbb{Q}_T}(f(X_0)) = \mathbb{E}_{\mathbb{P}}(f(X_0)\Lambda_T^{-1}) = \mathbb{E}_{\mathbb{P}}(f(X_0)\,\mathbb{E}_{\mathbb{P}}(\Lambda_T^{-1}|\mathcal{F}_0)) = \mathbb{E}_{\mathbb{P}}(f(X_0))$ for every bounded measurable f, which establishes the claim. Finally, to show that $\mathbb{P} \ll \mathbb{Q}_T$, note that Λ_T is (trivially) the corresponding Radon-Nikodym derivative. We are done. □

The following corollary follows trivially from the Bayes formula.

Corollary 7.1.7. *If $\mathbb{E}_{\mathbb{P}}(|f(X_0)|) < \infty$, then the filtered estimate is given by*

$$\pi_t(f) \equiv \mathbb{E}_{\mathbb{P}}(f(X_0)|\mathcal{F}_t^Y) = \frac{\mathbb{E}_{\mathbb{Q}_t}(f(X_0)\Lambda_t|\mathcal{F}_t^Y)}{\mathbb{E}_{\mathbb{Q}_t}(\Lambda_t|\mathcal{F}_t^Y)}.$$

As Y_t and X_0 are independent under $\mathbb{Q}_t$, one would expect that conditional expectations under $\mathbb{Q}_t$ can again be replaced by straightforward integration. The corresponding argument is essentially identical to the one used previously.

Lemma 7.1.8. *If $\mathbb{E}_{\mathbb{P}}(|f(X_0)|) < \infty$, then*

$$\sigma_t(f) \equiv \mathbb{E}_{\mathbb{Q}_t}(f(X_0)\Lambda_t|\mathcal{F}_t^Y) = \int_{\mathbb{R}} f(x)\, \exp\left(\kappa^{-2} x\, Y_t - \frac{1}{2}\kappa^{-2} x^2\, t\right) \mu_{X_0}(dx),$$

where μ_{X_0} is the law of the random variable X_0, and $\pi_t(f) = \sigma_t(f)/\sigma_t(1)$.

Proof. It suffices to check that $\mathbb{E}_{\mathbb{Q}_t}(I_A\, \mathbb{E}_{\mathbb{Q}_t}(f(X_0)\Lambda_t|\mathcal{F}_t^Y)) = \mathbb{E}_{\mathbb{Q}_t}(I_A f(X_0)\Lambda_t)$ for every $A \in \mathcal{F}_t^Y$. But this follows by an argument identical to the one employed in lemma 3.1.9. □

This is all we need to treat some specific examples!

Example 7.1.9 (Finite state case). Suppose that X_0 takes the values $x_1, \ldots, x_n$ with probabilities $p_1, \ldots, p_n$. Then we can write, for any function f,

$$\pi_t(f) = \frac{\sum_{i=1}^n p_i\, f(x_i)\, \exp(\kappa^{-2} x_i Y_t - \frac{1}{2}\kappa^{-2} x_i^2 t)}{\sum_{i=1}^n p_i\, \exp(\kappa^{-2} x_i Y_t - \frac{1}{2}\kappa^{-2} x_i^2 t)}.$$

Example 7.1.10 (Gaussian case). For Gaussian X_0 with mean μ and variance σ^2,

$$\sigma_t(f) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x)\, e^{\kappa^{-2} x Y_t - \kappa^{-2} x^2 t/2} e^{-(x-\mu)^2/2\sigma^2}\, dx.$$

This expression can be evaluated explicitly for $f(x) = x$ and $f(x) = x^2$, for example. The calculation is a little tedious, but gives the following answer:

$$\mathbb{E}_{\mathbb{P}}(X_0|\mathcal{F}_t^Y) = \frac{\kappa^2\mu + \sigma^2 Y_t}{\kappa^2 + \sigma^2 t}, \qquad \mathbb{E}_{\mathbb{P}}((X_0)^2|\mathcal{F}_t^Y) - (\mathbb{E}_{\mathbb{P}}(X_0|\mathcal{F}_t^Y))^2 = \frac{\kappa^2\sigma^2}{\kappa^2 + \sigma^2 t}.$$

Remark 7.1.11. Evidently, in the current setting (regardless of the law of X_0), the optimal estimate $\pi_t(f)$ depends on the observation history only through the random variable Y_t and in an explicitly computable fashion. This is an artefact, however, of this particularly simple model; in most cases the optimal estimate has a complicated dependence on the observation history, so that working directly with the Bayes formula, as we have done here, is not as fruitful (in general the Bayes formula does not lend itself to explicit computation). Instead, we will use the Bayes formula to obtain a stochatic differential equation for $\pi_t(f)$, which can subsequently be implemented *recursively* using, e.g., an Euler-Maruyama type method (at least in theory).

7.2 Nonlinear filtering for stochastic differential equations

The filtering problem

We work on $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, \mathbb{P})$. Consider a signal process of the form

$$X_t = X_0 + \int_0^t b(s, X_s, u_s)\, ds + \int_0^t \sigma(s, X_s, u_s)\, dW_s,$$

i.e., the signal which we would like to observe is the solution of a (possibly controlled) n-dimensional stochastic differential equation driven by the m-dimensional $\mathcal{F}_t$-Wiener process W_t. However, we do not have direct access to this signal; instead, we can only see the measurements taken by a noisy sensor, whose output is given by

$$Y_t = \int_0^t h(s, X_s, u_s)\, ds + \int_0^t K(s)\, dB_s,$$

where B_t is a p-dimensional $\mathcal{F}_t$-Wiener process independent of W_t. This model is of the "signal plus white noise" type: we observe essentially $y_t = h(t, u_t, X_t) + K(t)\,\xi_t$,

where ξ_t is white noise, but we work with the integrated form to obtain a sensible mathematical model (see the Introduction and section 3.3 for further discussion). In the equations above $b : [0,\infty[\times \mathbb{R}^n \times \mathbb{U} \to \mathbb{R}^n$, $\sigma : [0,\infty[\times \mathbb{R}^n \times \mathbb{U} \to \mathbb{R}^{n\times m}$, $h : [0,\infty[\times \mathbb{R}^n \times \mathbb{U} \to \mathbb{R}^p$, and $K : [0,\infty[\to \mathbb{R}^{p\times p}$ are measurable maps, $\mathbb{U} \subset \mathbb{R}^q$, and u_t is presumed to be adapted to the *observation filtration* $\mathcal{F}_t^Y = \sigma\{Y_s : s \le t\}$.

Remark 7.2.1. This is a rather general model; one often does not need this full-blown scenario! On the other hand, we are anticipating applications in control, so we have already included a control input. Note, however, that *the control may only depend (causally) on the observation process*; we can no longer use the state of the system X_t to determine u_t! This rules out the control strategies developed in the previous chapter, and we must reconsider our control problems in this more complicated setting.

The goal of the filtering problem is to compute, on the basis of the observations $\{Y_s : s \le t\}$, the (least mean square) optimal estimates $\pi_t(f) \equiv \mathbb{E}_{\mathbb{P}}(f(X_t)|\mathcal{F}_t^Y)$, at least for a sufficiently large class of functions f. To keep matters as simple as possible we will be content to operate under rather stringent technical conditions:

1. The equation for (X_t, Y_t) has a unique $\mathcal{F}_t$-adapted solution;
2. $K(t)$ is invertible for all t and $K(t), K(t)^{-1}$ are locally bounded;
3. b, σ, h are bounded functions.

The last condition is particularly restrictive, and can be weakened significantly (see section 7.7 for detailed references). This tends to become a rather technical exercise. By restricting ourselves to bounded coefficients, we will be able to concentrate on the essential ideas without being bogged down by a large number of technicalities.

Unfortunately, one of the most important examples of the theory, the Kalman-Bucy filter, does not satisfy condition 3. We will circumvent the technicalities by treating this case separately, using a different approach, in the next section.

The Kallianpur-Striebel formula

Our first order of business is essentially to repeat example 7.1.1 in the current setting; that is, we will find an explicit representation of the filtered estimates $\pi_t(f)$ in terms of a particularly convenient reference measure $\mathbb{Q}$. How should we choose this measure? The discussion in the previous section suggests that we should try to choose $\mathbb{Q}$ such that X_t and $\mathcal{F}_t^Y$ are independent under $\mathbb{Q}$. The presence of the control u_t makes this difficult, but fortunately it will suffice to make X_0, W_t and Y_t independent.

Lemma 7.2.2. *Define the measure $\mathbb{Q}_T \ll \mathbb{P}$ by setting*

$$\frac{d\mathbb{Q}_T}{d\mathbb{P}} = \exp\left[-\int_0^T (K(t)^{-1}h(t,X_t,u_t))^* dB_t - \frac{1}{2}\int_0^T \|K(t)^{-1}h(t,X_t,u_t)\|^2 dt\right].$$

Then under $\mathbb{Q}_T$, the process $(W_t, \bar{Y}_t)_{t\in[0,T]}$ is an $(m+p)$-dimensional $\mathcal{F}_t$-Wiener process (so, in particular, independent of X_0), where

$$\bar{Y}_t = \int_0^t K(s)^{-1}h(s,X_s,u_s)\,ds + B_t = \int_0^t K(s)^{-1}\,dY_s.$$

Moreover, $\mathbb{P} \ll \mathbb{Q}_T$ *with*

$$\frac{d\mathbb{P}}{d\mathbb{Q}_T} = \exp\left[\int_0^T (K(t)^{-1}h(t,X_t,u_t))^* d\bar{Y}_t - \frac{1}{2}\int_0^T \|K(t)^{-1}h(t,X_t,u_t)\|^2\,dt\right].$$

Remark 7.2.3. In filtering theory, we will regularly encounter stochastic intergrals with respect to processes such as $Y_t, \bar{Y}_t$. As $\bar{Y}_t$ is a Wiener process under $\mathbb{Q}_T$, we can construct these integrals under $\mathbb{Q}_T$ by our usual approach. We then indeed find, e.g.,

$$\int_0^t F_s^* d\bar{Y}_s = \int_0^t F_s^* K(s)^{-1} h(s,X_s,u_s)\,ds + \int_0^t F_s^* dB_s,$$

as one would naively think, by virtue of the fact that $\mathbb{Q}_T$ and $\mathbb{P}$ are mutually absolutely continuous; this is easily verified when F_t is simple and bounded, and can be extended to any integrable F_t through the usual process of taking limits and localization. Alternatively, one can set up a more general integration theory that is not restricted to Wiener process integrands, so that these integrals can be constructed under both measures. As this is an introductory course, we will not dwell on these technical issues; we will be content to accept the fact that stochastic integrals are well-behaved under absolutely continuous changes of measure (see, e.g., [Pro04] or [RY99]).

Proof. By conditions 2 and 3, Novikov's condition is satisfied. We can thus apply Girsanov's theorem to the $(m+p)$-dimensional process $(dW_t, d\bar{Y}_t) = d\mathbf{Y}_t = \mathbf{H}_t\,dt + d\mathbf{W}_t$, where $\mathbf{W}_t = (W_t, B_t)$ and $\mathbf{H}_t = (0, K(t)^{-1}(t,X_t,u_t))$. We find that under $\mathbb{Q}_T$, the process $\{\mathbf{Y}_t\}_{t\in[0,T]}$ is an $\mathcal{F}_t$-Wiener process; in particular, W_t and $\bar{Y}_t$ are independent Wiener processes and both are independent of X_0 (as X_0 is $\mathcal{F}_0$-measurable, and these are $\mathcal{F}_t$-Wiener processes).

Now note that $\mathbb{P} \ll \mathbb{Q}_T$ follows immediately from the fact that $d\mathbb{Q}_T/d\mathbb{P}$ is strictly positive, where $d\mathbb{P}/d\mathbb{Q}_T = (d\mathbb{Q}_T/d\mathbb{P})^{-1}$ (which is precisely the expression in the statement of the lemma). After all, $\mathbb{E}_{\mathbb{Q}_T}(Z\,(d\mathbb{Q}_T/d\mathbb{P})^{-1}) = \mathbb{E}_{\mathbb{P}}(Z\,d\mathbb{Q}_T/d\mathbb{P}\,(d\mathbb{Q}_T/d\mathbb{P})^{-1}) = \mathbb{E}_{\mathbb{P}}(Z)$. □

Applying the Bayes formula is now straightforward. We will frequently write

$$\Lambda_t = \exp\left[\int_0^T (K(t)^{-1}h(t,X_t,u_t))^* d\bar{Y}_t - \frac{1}{2}\int_0^T \|K(t)^{-1}h(t,X_t,u_t)\|^2\,dt\right],$$

which is a martingale under $\mathbb{Q}_T$ for $t \le T$ (Novikov). The Bayes formula now gives:

Corollary 7.2.4. *If* $\mathbb{E}_{\mathbb{P}}(|f(X_t)|) < \infty$*, the filtered estimate* $\pi_t(f)$ *is given by*

$$\pi_t(f) = \mathbb{E}_{\mathbb{P}}(f(X_t)|\mathcal{F}_t^Y) = \frac{\mathbb{E}_{\mathbb{Q}_t}(f(X_t)\Lambda_t|\mathcal{F}_t^Y)}{\mathbb{E}_{\mathbb{Q}_t}(\Lambda_t|\mathcal{F}_t^Y)} = \frac{\sigma_t(f)}{\sigma_t(1)},$$

where we have defined the unnormalized estimate $\sigma_t(f) = \mathbb{E}_{\mathbb{Q}_t}(f(X_t)\Lambda_t|\mathcal{F}_t^Y)$.

This expression is called the *Kallianpur-Striebel formula*. In fact, Kallianpur and Striebel went a little further: they actually expressed the unnormalized conditional expectation $\sigma_t(f)$ as an integral over a part of the probability space, just like we did in the previous section, thus making the analogy complete (see, e.g., [LS01a, section 7.9], for the relevant argument). However, we will find is just as easy to work directly with the conditional expectations, so we will not bother to make this extra step.

Remark 7.2.5. Note that Girsanov's theorem implies $\mathbb{Q}_t(A) = \mathbb{Q}_T(A)$ for any $A \in \mathcal{F}_t$; in particular, $\mathbb{E}_{\mathbb{Q}_t}(f(X_t)\Lambda_t|\mathcal{F}_t^Y) = \mathbb{E}_{\mathbb{Q}_T}(f(X_t)\Lambda_t|\mathcal{F}_t^Y) = \mathbb{E}_{\mathbb{Q}_T}(f(X_t)\Lambda_T|\mathcal{F}_t^Y)$ (why?). We will occasionally use this, e.g., in the proof proposition 7.2.6 below.

What progress have we made? Quite a lot, as a matter of fact, though it is not immediately visible. What we have gained by representing $\pi_t(f)$ in this way is that the filtering problem is now expressed in terms of a particularly convenient measure. To proceed, we can turn the crank on our standard machinery: the Itô rule *et al.*

The Zakai equation

For the time being, we will concentrate not on $\pi_t(f)$, but on its unnormalized counterpart $\sigma_t(f)$. Our next order of business is to find an explicit expression for $\sigma_t(f)$. This is not too difficult; we simply apply Itô's rule to the process $f(X_t)\Lambda_t$, then try to compute the conditional expectation of this expression (making use of the independence properties under $\mathbb{Q}_t$). Once we have accomplished this, the remainder is easy: another application of the Itô rule gives an expression for the normalized quantity $\pi_t(f)$.

Proposition 7.2.6 (Zakai equation). *Let f be C^2 and suppose that f and all its derivatives are bounded. Then we can write*

$$\sigma_t(f) = \sigma_0(f) + \int_0^t \sigma_s(\mathscr{L}_s^u f)\,ds + \int_0^t \sigma_s(K(s)^{-1}h_s^u f)^* d\bar{Y}_s,$$

where $\sigma_0(f) = \mathbb{E}_{\mathbb{P}}(f(X_0))$ and $(h_s^u f)(x) = h(s,x,u_s)f(x)$.

Proof. Using Itô's rule, we find that

$$f(X_t)\Lambda_t = f(X_0) + \int_0^t \Lambda_s\,\mathscr{L}_s^u f(X_s)\,ds + \int_0^t \Lambda_s\,(\nabla f(X_s))^*\sigma(s,X_s,u_s)\,dW_s + \int_0^t \Lambda_s\,f(X_s)\,(K(s)^{-1}h(s,X_s,u_s))^* d\bar{Y}_s.$$

By our boundedness assumptions, all the integrands are in $\mathcal{L}^2(\mu_t \times \mathbb{Q}_t)$. Hence we can compute

$$\mathbb{E}_{\mathbb{Q}_t}(f(X_t)\Lambda_t|\mathcal{F}_t^Y) = \mathbb{E}_{\mathbb{Q}_t}(f(X_0)|\mathcal{F}_t^Y) + \int_0^t \mathbb{E}_{\mathbb{Q}_t}(\Lambda_s\,\mathscr{L}_s^u f(X_s)|\mathcal{F}_s^Y)\,ds + \int_0^t \mathbb{E}_{\mathbb{Q}_t}(\Lambda_s\,K(s)^{-1}h(s,X_s,u_s)\,f(X_s)|\mathcal{F}_s^Y)^* d\bar{Y}_s,$$

where we have used lemma 7.2.7 below. It remains to note that X_0 and $\mathcal{F}_t^Y$ are independent under $\mathbb{Q}_t$, so $\mathbb{E}_{\mathbb{Q}_t}(f(X_0)|\mathcal{F}_t^Y) = \mathbb{E}_{\mathbb{Q}_t}(f(X_0)) = \mathbb{E}_{\mathbb{P}}(f(X_0))$. □

We have used the following elementary result.

Lemma 7.2.7. *Let W_t, V_t be independent $\mathcal{F}_t$-Wiener processes, let $F_\cdot \in \mathcal{L}^2(\mu_t \times \mathbb{P})$ be $\mathcal{F}_t$-adapted, and define the sub-filtration $\mathcal{F}_t^W = \sigma\{W_s : s \le t\} \subset \mathcal{F}_t$. Then*

$$\mathbb{E}\left[\int_0^t F_s\,dW_s \,\middle|\, \mathcal{F}_t^W\right] = \int_0^t \mathbb{E}(F_s|\mathcal{F}_s^W)\,dW_s, \qquad \mathbb{E}\left[\int_0^t F_s\,dV_s \,\middle|\, \mathcal{F}_t^W\right] = 0.$$

Moreover, a similar result holds for the time integral:

$$\mathbb{E}\left[\int_0^t F_s\,ds\,\middle|\,\mathcal{F}_t^W\right] = \int_0^t \mathbb{E}(F_s|\mathcal{F}_s^W)\,ds.$$

Proof. Choose any $A \in \mathcal{F}_t^W$, and note that by the Itô representation theorem

$$I_A = \mathbb{P}(A) + \int_0^t H_s\,dW_s$$

for some $\mathcal{F}_t^W$-adapted process $H_\cdot \in \mathcal{L}^2(\mu_t \times \mathbb{P})$. Let us now apply the polarization identity $2\,\mathbb{E}(I_t(H)I_t(F)) = \mathbb{E}((I_t(H)+I_t(F))^2) - \mathbb{E}(I_t(H)^2) - \mathbb{E}(I_t(F)^2)$ and the Itô isometry:

$$\mathbb{E}\left[I_A\int_0^t F_s\,dW_s\right] = \mathbb{E}\left[\int_0^t F_s\,H_s\,ds\right] = \mathbb{E}\left[\int_0^t \mathbb{E}(F_s|\mathcal{F}_s^W)\,H_s\,ds\right],$$

where we have used Fubini's theorem and the tower property of the conditional exepctation in the last step. But by applying the same steps with F_s replaced by $\mathbb{E}(F_s|\mathcal{F}_s^W)$, we find that

$$\mathbb{E}\left[I_A\int_0^t F_s\,dW_s\right] = \mathbb{E}\left[I_A\int_0^t \mathbb{E}(F_s|\mathcal{F}_s^W)\,dW_s\right] \quad \text{for all} \quad A \in \mathcal{F}_t^W.$$

Hence the first statement follows by the Kolmogorov definition of the conditional expectation. The second statement follows in the same way. To establish the last statmement, note that

$$\mathbb{E}\left[I_A\int_0^t F_s\,ds\right] = \mathbb{E}\left[I_A\int_0^t \mathbb{E}(F_s|\mathcal{F}_t^W)\,ds\right] = \mathbb{E}\left[I_A\int_0^t \mathbb{E}(F_s|\mathcal{F}_s^W)\,ds\right]$$

for $A \in \mathcal{F}_t^W$, where the first equality follows by using Fubini's theorem and the tower property of the conditional expectation, and the second equality follows as F_s, being $\mathcal{F}_s$-measurable, is independent of $\mathcal{F}_{t,s}^W = \sigma\{W_r - W_s : s \le r \le t\}$, and $\mathcal{F}_t^W = \sigma\{\mathcal{F}_s^W, \mathcal{F}_{t,s}^W\}$. □

The Kushner-Stratonovich equation and the innovations process

Now that we have an equation for $\sigma_t(f)$, the equation for $\pi_t(f)$ is simply a matter of applying Itô's rule. Let us see what happens.

Proposition 7.2.8 (Kushner-Stratonovich equation). *Let f be C^2 and suppose that f and all its derivatives are bounded. Then we can write*

$$\pi_t(f) = \pi_0(f) + \int_0^t \pi_s(\mathscr{L}_s^u f)\,ds + \int_0^t \{\pi_s(K(s)^{-1}h_s^u f) - \pi_s(f)\,\pi_s(K(s)^{-1}h_s^u)\}^*(d\bar{Y}_s - \pi_s(K(s)^{-1}h_s^u)\,ds),$$

where $\pi_0(f) = \mathbb{E}_{\mathbb{P}}(f(X_0))$ and $(h_s^u f)(x) = h(s,x,u_s)f(x)$.

Proof. $\sigma_t(1)$ is strictly positive $\mathbb{P}$-a.s. (by lemma 7.1.3), and hence also $\mathbb{Q}_T$-a.s. ($t \le T$) as $\mathbb{Q}_T \ll \mathbb{P}$. Hence we can apply Itô's rule to compute $\sigma_t(f)(\sigma_t(1))^{-1}$. Straightforward computations and application of the Kallianpur-Striebel formula yield the desired expression. □

In the Kushner-Stratonovich equation, the interesting process

$$\bar{B}_t = \bar{Y}_t - \int_0^t \pi_s(K(s)^{-1} h_s^u)\, ds$$

just popped up while applying Itô's rule; $\bar{B}_t$ is called the *innovations process*. It has an important property that can be extremely useful in control applications.

Proposition 7.2.9. *Under $\mathbb{P}$, the innovation $\bar{B}_t$ is an $\mathcal{F}_t^Y$-Wiener process, so we have*

$$\pi_t(f) = \pi_0(f) + \int_0^t \pi_s(\mathscr{L}_s^u f)\, ds + \int_0^t \{K(s)^{-1}(\pi_s(h_s^u f) - \pi_s(f)\,\pi_s(h_s^u))\}^* d\bar{B}_s.$$

Proof. $\bar{B}_t$ is clearly $\mathcal{F}_t^Y$-adapted and has continuous sample paths. To prove that it is a Wiener process, we proceed essentially as in the proof of Girsanov's theorem: we will show that

$$\mathbb{E}_{\mathbb{P}}(e^{i\alpha^*(\bar{B}_t - \bar{B}_s) + i\beta Z}) = e^{-\|\alpha\|^2 (t-s)/2}\, \mathbb{E}_{\mathbb{P}}(e^{i\beta Z}) \quad \text{for any } \mathcal{F}_s^Y\text{-measurable } Z.$$

It suffices to prove that $\mathbb{E}_{\mathbb{P}}(e^{i\alpha^*(\bar{B}_t - \bar{B}_s)} | \mathcal{F}_s^Y) = e^{-\|\alpha\|^2 (t-s)/2}$, as this statement then follows.

To proceed, let us apply Itô's rule to $e^{i\alpha^* \bar{B}_t}$. This gives

$$\begin{aligned} e^{i\alpha^* \bar{B}_t} = e^{i\alpha^* \bar{B}_s} &+ \int_s^t e^{i\alpha^* \bar{B}_r}\, i\alpha^* dB_r \\ &+ \int_s^t e^{i\alpha^* \bar{B}_r} \left[i\alpha^* K(r)^{-1}(h(r, X_r, u_r) - \pi_r(h_r^u)) - \frac{\|\alpha\|^2}{2} \right] dr. \end{aligned}$$

We now condition on $\mathcal{F}_s^Y$. We claim that the stochastic integral vanishes; indeed, it is an $\mathcal{F}_t$-martingale, so vanishes when conditioned on $\mathcal{F}_s$, and $\mathcal{F}_s^Y \subset \mathcal{F}_s$ establishes the claim. Moreover, note that $\mathbb{E}_{\mathbb{P}}(e^{i\alpha^* \bar{B}_r}\, \mathbb{E}_{\mathbb{P}}(h(r, X_r, u_r) | \mathcal{F}_r^Y) | \mathcal{F}_s^Y) = \mathbb{E}_{\mathbb{P}}(e^{i\alpha^* \bar{B}_r} h(r, X_r, u_r) | \mathcal{F}_s^Y)$, as $e^{i\alpha^* \bar{B}_r}$ is $\mathcal{F}_r^Y$-measurable and $\mathcal{F}_s^Y \subset \mathcal{F}_r^Y$. Hence as in the proof of lemma 7.2.7, we find

$$\mathbb{E}_{\mathbb{P}}(e^{i\alpha^* \bar{B}_t} | \mathcal{F}_s^Y) = e^{i\alpha^* \bar{B}_s} - \frac{\|\alpha\|^2}{2} \int_s^t \mathbb{E}_{\mathbb{P}}(e^{i\alpha^* \bar{B}_r} | \mathcal{F}_s^Y)\, dr.$$

But this equation has the unique solution $\mathbb{E}_{\mathbb{P}}(e^{i\alpha^* \bar{B}_t} | \mathcal{F}_s^Y) = e^{i\alpha^* \bar{B}_s - \|\alpha\|^2 (t-s)/2}$. □

We will postpone providing an application of this result until section 7.5.

How to compute the filtered estimates?

We have now obtained a stochastic integral expression—the Kushner-Stratonovich equation—for $\pi_t(f)$. However, as you have most likely already noticed, this is not a stochastic differential equation for $\pi_t(f)$. After all, the integrands in this equation depend on $\pi_t(\mathscr{L}_t^u f)$, $\pi_t(h_s^u f)$, etc., which can not be expressed (in general) as functions of $\pi_t(f)$! Hence this equation can not be used by itself to compute $\pi_t(f)$.

You might hope that if we choose the functions f correctly, then the filtering equations would *close*. For example, suppose there is a collection $f_1, \ldots, f_n$, for which we can show that $\pi_t(\mathscr{L}_t^u f_i)$, $\pi_t(h_t^u f_i)$ and $\pi_t(h_t^u)$ can again be expressed as functions of $\pi_t(f_1), \ldots, \pi_t(f_n)$. The Kushner-Stratonovich equation for $(\pi_t(f_1), \ldots, \pi_t(f_n))$

would then reduce to an SDE which can be computed, e.g., using the Euler-Maruyama method. Unfortunately, it turns out that this is almost never the case—in most cases *no finite-dimensional realization of the filtering equation exists*. There is one extremely important exception to this rule: when b, h are linear, σ is constant and X_0 is a Gaussian random variable, we obtain the finite-dimensional *Kalman-Bucy filter* which is very widely used in applications; this is the topic of the next section. However, a systematic search for other finite-dimensional filters has unearthed few examples of practical relevance (see, e.g., [HW81, Mit82, Par91, HC99]).[1]

Of course, it would be rather naive to expect that the conditional expectations $\mathbb{E}_{\mathbb{P}}(f(X_t)|\mathcal{F}_t^Y)$ can be computed in a finite-dimensional fashion. After all, in most cases, even the *unconditional* expectation $\mathbb{E}_{\mathbb{P}}(f(X_t))$ can not be computed by solving a finite-dimensional equation! Indeed, the Itô rule gives

$$\frac{d}{dt}\,\mathbb{E}_{\mathbb{P}}(f(X_t)) = \mathbb{E}_{\mathbb{P}}(\mathscr{L}_t^u f(X_t)),$$

which depends on $\mathbb{E}_{\mathbb{P}}(\mathscr{L}_t^u f(X_t))$; and the equation for $\mathbb{E}_{\mathbb{P}}(\mathscr{L}_t^u f(X_t))$ will depend on $\mathbb{E}_{\mathbb{P}}(\mathscr{L}_t^u \mathscr{L}_t^u f(X_t))$ (if b, σ, f are sufficiently smooth), etc., so that we will almost certainly not obtain a closed set of equations for any collection of functions $f_1, \ldots, f_n$. (Convince yourself that the case where b, f are linear and σ is constant is an exception!) To actually compute $\mathbb{E}_{\mathbb{P}}(f(X_t))$ (in the absence of control, for example), we have two options: either we proceed in Monte Carlo fashion by averaging a large number of simulated (random) sample paths of X_t, or we solve one of the PDEs associated with the SDE for X_t: the Kolmogorov forward or backward equations. The latter are clearly infinite-dimensional, while in the former case we would need to average an infinite number of random samples to obtain an exact answer for $\mathbb{E}_{\mathbb{P}}(f(X_t))$.

We are faced with a similar choice in the filtering problem. If we are not in the Kalman-Bucy setting, or one which is sufficiently close that we are willing to linearize our filtering model (the latter gives rise to the so-called *extended Kalman filter* [Par91]), we will have to find some numerically tractable approximation. One popular approach is of the Monte Carlo type; the so-called *particle filtering methods*, roughly speaking, propagate a collection of random samples in such a way that the probability of observing these "particles" in a certain set A is an approximation of $\pi_t(I_A)$ (or of a related object). Particle methods are quite effective and are often used, e.g., in tracking, navigation, and robotics applications. Unfortunately the details of such methods are beyond our scope, but see [Del04, CL97] for discussion and further references.

Another approach is through PDEs. For simplicity, let us consider (on a formal level) the filtering counterpart of the Kolmogorov forward equation. We will assume that the filtering problem possesses a *conditional density*, i.e., that there is a *random* density $p_t(x)$, which is only a functional of the observations $\mathcal{F}_t^Y$, such that

$$\pi_t(f) = \mathbb{E}_{\mathbb{P}}(f(X_t)|\mathcal{F}_t^Y) = \int f(x)\, p_t(x)\, dx.$$

[1] An important class of finite-dimensional nonlinear filters with applications, e.g., in speech recognition, are those for which the signal X_t is not the solution of a stochastic differential equation, as in this section, but a finite-state Markov process [Won65, LS01a]. We will discuss a special case in section 7.4.

Formally integrating by parts in the Kushner-Stratonovich equation, we obtain

$$dp_t(x) = (\mathscr{L}_t^u)^* p_t(x)\,dt + \\ p_t(x)\{K(t)^{-1}(h(t,x,u_t) - \pi_t(h_t^u))\}^*\,(d\bar{Y}_t - \pi_t(K(t)^{-1}h_t^u)\,dt),$$

which is a nonlinear stochastic partial integro-differential equation. It is not an easy task to make mathematical sense of this equation; how should the equation even be interpreted, and do such equations have solutions? For details on such questions see, e.g., [Kun90]. If we wish to work with PDEs, however, it usually makes more sense to work with the Zakai equation instead. Assuming the existence of $q_t(x)$ such that

$$\sigma_t(f) = \int f(x)\,q_t(x)\,dx, \qquad p_t(x) = \frac{q_t(x)}{\int q_t(x)\,dx},$$

we can formally obtain the Zakai equation in PDE form:

$$dq_t(x) = (\mathscr{L}_t^u)^* q_t(x)\,dt + q_t(x)\,(K(t)^{-1}h(t,x,u_t))^* d\bar{Y}_t.$$

At least this equation is a linear stochastic partial differential equation, a much more well-posed object. It is still much too difficult for us, but the corresponding theory can be found, e.g., in [Par82, Par91, Ben92, Kun90]. The Zakai PDE can now be the starting point for further approximations, e.g., Galerkin-type methods [GP84], spectral methods [LMR97], or projection onto a finite-dimensional manifold [BHL99].

Finally, there is a third approach which is similar to the method that we have already encountered in the control setting. We can approximate our signal process by a discrete time finite-state Markov process, and introduce an appropriate approximation to the observation process; this can be done, e.g., by introducing a suitable finite-difference approximation, as we did in the last chapter. The optimal filter for the approximate signal and observations is a finite-dimensional recursion, which can be shown to converge to the solution of the optimal filtering problem [Kus77, KD01].

For a recent review on numerical methods in nonlinear filtering, see [Cri02].

Example 7.2.10. For sake of example, and as we already have some experience with such approximations, let us discuss an extremely simple Markov chain approximation for a nonlinear filtering problem. This is not necessarily the method of choice for such a problem, but will serve as a simple demonstration.

Consider a signal process θ_t on the circle which satisfies

$$d\theta_t = \omega\,dt + \nu\,dW_t \qquad (\mathrm{mod}\ 2\pi),$$

where θ_0 is uniformly distributed on the circle. We consider the observations process

$$dY_t = \sin(\theta_t)\,dt + \kappa\,dB_t,$$

and our goal is to estimate θ_t given $\mathcal{F}_t^Y$. Such a model can be used in phase tracking problems (e.g., in a phase lock loop), where the goal is to estimate the drifting phase of an oscillating signal (with carrier frequency ω) from noisy observations [Wil74].

To proceed, it is easiest to first approximate the signal process θ_t by a Markov chain, so that we can subsequently formulate a filtering problem for this Markov chain. To this end, let us consider the Kolmogorov backward equation for θ_t:

$$\frac{\partial\, u(t,\theta)}{\partial t} = \frac{\nu^2}{2}\frac{\partial^2 u(t,\theta)}{\partial \theta^2} + \omega\,\frac{\partial\, u(t,\theta)}{\partial \theta}, \qquad u(0,x) = f(x),$$

so that $u(t, X_0) = \mathbb{E}(f(X_t)|X_0)$. Substituting our usual finite-difference approximations on the right, and using a forward difference for the time derivative, we obtain

$$u(t+\Delta,\theta) \approx \left[1 - \frac{\Delta\nu^2}{\delta^2}\right]u(t,\theta) + \frac{\Delta}{2\delta}\left[\frac{\nu^2}{\delta} + \omega\right]u(t,\theta+\delta) + \frac{\Delta}{2\delta}\left[\frac{\nu^2}{\delta} - \omega\right]u(t,\theta-\delta),$$

where Δ is the time step size, $\delta = \pi/N$ is the discretization step on the circle, i.e., we discretize $]0, 2\pi]$ into $S_\delta = \{k\delta : k = 1, \ldots, 2N\}$, and circular boundary conditions are implied. But this expression is easily seen to be the Kolmogorov backward equation for the Markov chain x_n with values in S_δ and with transition probabilities

$$\mathbb{P}(x_n = k\delta | x_{n-1} = k\delta) = 1 - \frac{\Delta\nu^2}{\delta^2}, \quad \mathbb{P}(x_n = (k\pm1)\delta | x_{n-1} = k\delta) = \frac{\Delta}{2\delta}\left[\frac{\nu^2}{\delta} \pm \omega\right],$$

provided that Δ, δ are sufficiently small that these values are in the interval $[0, 1]$.

Now that we have obtained our approximate Markov chain, how can we use this to approximate the optimal filter? Rather than using the Zakai or Kushner-Stratonovich equations, let us use directly the Kallianpur-Striebel formula. In the current case, we can write the optimal filter as $\pi_t(f) = \sigma_t(f)/\sigma_t(1)$, where

$$\sigma_t(f) = \mathbb{E}_{\mathbb{Q}_t}\left[f(X_t)\,\exp\left(\int_0^t \kappa^{-2}\sin(\theta_t)\,dY_t - \frac{1}{2}\int_0^t \kappa^{-2}\sin^2(\theta_t)\,dt\right)\middle|\,\mathcal{F}_t^Y\right].$$

To approximate this expression, we replace θ_t by the approximate Markov chain x_n, and we replace the integrals by Euler-Maruyama type sums. This gives

$$\sigma_{n\Delta}(f) \approx \mathbb{E}_{\mathbb{Q}}\left[f(x_n)\,e^{\kappa^{-2}\sum_{m=0}^{n-1}\{\sin(x_m)\,(Y_{(m+1)\Delta} - Y_{m\Delta}) - \frac{1}{2}\sin^2(x_m)\,\Delta\}}\middle|\,\mathcal{F}_{n\Delta}^Y\right],$$

where $\mathbb{Q}$ is the measure under which x_n and $\mathcal{F}_t^Y$ are independent. A weak convergence argument [KD01, Kus77] guarantees that this approximate expression does indeed converge, as $\Delta, \delta \to 0$, to the exact unnormalized estimate $\sigma_{n\Delta}(f)$.

Remark 7.2.11. Consider the following discrete time filtering problem: the signal is our Markov chain x_n, while at time n we observe $y_n = \sin(x_n)\,\Delta + \kappa\,\xi_n$, where ξ_n is a Gaussian random variable with mean zero and variance Δ, independent of the signal process. Using the Bayes formula, you can easily verify that

$$\mathbb{E}(f(x_n)|\mathcal{F}_n^y) = \frac{\mathbb{E}_{\mathbb{Q}}(f(x_n)\Lambda_n|\mathcal{F}_n^y)}{\mathbb{E}_{\mathbb{Q}}(\Lambda_n|\mathcal{F}_n^y)}, \quad \Lambda_n = e^{\kappa^{-2}\sum_{m=0}^{n-1}\{\sin(x_m)\,y_m - \frac{1}{2}\sin^2(x_m)\,\Delta\}},$$

where $\mathcal{F}_n^y = \sigma\{y_m : m \le n\}$ and $\mathbb{Q}$ is the measure under which $\{x_n\}$ and $\{y_n\}$ are independent. Evidently our approximate filter is again a filter for an approximate problem, just like the Markov chain approximations in the stochastic control setting.

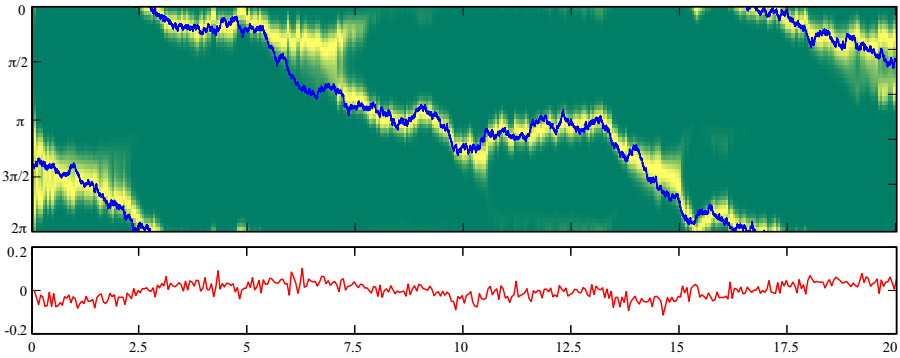

Figure 7.1. Numerical solution of example 7.2.10 with $\omega = \nu = .5$, $\kappa = .1$, $\Delta = .05$ and $N = 25$, on the interval $t \in [0, 20]$. The top plot shows θ_t (blue line) and the approximate conditional distribution $\tilde{\pi}_n$ (shaded background), while the bottom plot shows the observation increments $Y_{(m+1)\Delta} - Y_{m\Delta}$ used by the approximate filter (red). The blue and red plots were computed by the Euler-Maruyama method with a time step much smaller than Δ.

To compute the approximate filter recursively, note that

$$\tilde{\sigma}_n(f) = \mathbb{E}_{\mathbb{Q}}\left[\mathbb{E}_{\mathbb{Q}}(f(x_n)|x_{n-1})\, e^{\kappa^{-2}\sum_{m=0}^{n-1}\{\sin(x_m)\,(Y_{(m+1)\Delta}-Y_{m\Delta})-\frac{1}{2}\sin^2(x_m)\,\Delta\}}\,\middle|\,\mathcal{F}^Y_{n\Delta}\right],$$

where we have written $\tilde{\sigma}_n(f)$ for the approximate expression for $\sigma_{n\Delta}(f)$. To see this, it suffices to note that as Y_t is independent of x_n under $\mathbb{Q}$, and as x_n has the same law under $\mathbb{P}$ and $\mathbb{Q}$, we can write $\mathbb{E}_{\mathbb{Q}}(f(x_n)|\sigma\{\mathcal{F}^Y_{n\Delta}, \mathcal{F}^x_{n-1}\}) = \mathbb{E}_{\mathbb{P}}(f(x_n)|x_{n-1})$ using the Markov property (where $\mathcal{F}^x_n = \sigma\{x_m : m \le n\}$); the claim follows directly using the tower property of the conditional expectation. But then evidently

$$\tilde{\sigma}_n(f) = \tilde{\sigma}_{n-1}\left(\mathbb{E}_{\mathbb{P}}(f(x_n)|x_{n-1} = \cdot)\, e^{\kappa^{-2}\{\sin(\cdot)\,(Y_{(n)\Delta}-Y_{(n-1)\Delta})-\frac{1}{2}\sin^2(\cdot)\,\Delta\}}\right),$$

which is the discrete time analog of the Zakai equation. We can now turn this into a closed-form recursion as follows. Define $\tilde{\sigma}^k_n = \tilde{\sigma}_n(I_{\{k\delta\}})$, denote by P the matrix with elements $P_{k\ell} = \mathbb{P}(x_n = \ell\delta | x_{n-1} = k\delta)$, and by $\Lambda(y)$ the diagonal matrix with $(\Lambda(y))_{kk} = e^{\kappa^{-2}\{\sin(k\delta)\,y - \sin^2(k\delta)\,\Delta/2\}}$. Then you can easily verify that

$$\tilde{\sigma}_n = \Lambda(Y_{n\Delta} - Y_{(n-1)\Delta})P^*\tilde{\sigma}_{n-1}, \qquad \tilde{\pi}_n = \frac{\Lambda(Y_{n\Delta} - Y_{(n-1)\Delta})P^*\tilde{\pi}_{n-1}}{\sum_k(\Lambda(Y_{n\Delta} - Y_{(n-1)\Delta})P^*\tilde{\pi}_{n-1})^k},$$

where $\tilde{\pi}^k_n = \tilde{\sigma}_n(I_{\{k\delta\}})/\tilde{\sigma}_n(1)$ is the approximate conditional probability of finding θ_t in the kth discretization interval at time $n\Delta$, given the observations up to that time.

A numerical simulation is shown in figure 7.1; that the approximate filter provides good estimates of the signal location is evident. A curious effect should be pointed out: note that whenever the signal crosses either $\pi/2$ or $3\pi/2$, the conditional distribution briefly becomes bimodal. This is to be expected, as these are precisely the peaks

of the observation function $\sin(x)$; convince yourself that around these peaks, the filter cannot distinguish purely from the observations in which direction the signal is moving! This causes the conditional distribution to have "ghosts" which move in the opposite direction. However, the "ghosts" quickly dissipate away, as prolonged motion in the opposite direction is incompatible with the signal dynamics (of course, the effect is more pronounced if ω is close to zero). Thus the filter does its job in utilizing both the information gained from the observations and the known signal dynamics.

7.3 The Kalman-Bucy filter

In the previous section we considered the nonlinear filtering problem in a rather general setting. We have seen that this problem can be solved explicitly, but that implementation of the resulting equations is often computationally intensive. Nonetheless, this can be well worth the effort in a variety of applications (but may be too restrictive in others). On the other hand, in this and the next section we will discuss two important cases where the optimal filtering equations can be expressed in finite-dimensional closed form. These filters are consequently easily implemented in practice, and are found in a wide range of applications throughout science and engineering (particularly the Kalman-Bucy filter, which is the topic of this section).

The linear filtering problem

We will consider the case where b, h are linear and σ is constant, i.e.,

$$\begin{aligned} dX_t &= A(t)X_t\,dt + B(t)u_t\,dt + C(t)\,dW_t, \\ dY_t &= H(t)X_t\,dt + K(t)\,dB_t. \end{aligned}$$

Here $A(t)$, $B(t)$, $C(t)$, $H(t)$, and $K(t)$ are non-random matrices of dimensions $n\times n$, $n\times k$, $n\times m$, $p\times n$, and $p\times p$, respectively, and u_t is a k-dimensional control input which is presumed to be $\mathcal{F}_t^Y$-adapted. We will make the following assumptions:

1. X_0 is a Gaussian random variable with mean $\hat{X}_0$ and covariance $\hat{P}_0$;
2. The equation for (X_t, Y_t) has a unique $\mathcal{F}_t$-adapted solution;
3. $K(t)$ is invertible for all t;
4. $A(t), B(t), C(t), H(t), K(t), K(t)^{-1}$ are continuous.

The goal of the *linear filtering problem* is to compute the conditional mean $\hat{X}_t = \mathbb{E}(X_t|\mathcal{F}_t^Y)$ and error covariance $\hat{P}_t = \mathbb{E}((X_t-\hat{X}_t)(X_t-\hat{X}_t)^*)$. We will prove:

Theorem 7.3.1 (Kalman-Bucy). *Under suitable conditions on the control u_t,*

$$\begin{aligned} d\hat{X}_t &= A(t)\hat{X}_t\,dt + B(t)u_t\,dt + \hat{P}_t(K(t)^{-1}H(t))^* d\bar{B}_t, \\ \frac{d\hat{P}_t}{dt} &= A(t)\hat{P}_t + \hat{P}_t A(t)^* - \hat{P}_t H(t)^*(K(t)K(t)^*)^{-1}H(t)\hat{P}_t + C(t)C(t)^*, \end{aligned}$$

where $d\bar{B}_t = K(t)^{-1}(dY_t - H(t)\hat{X}_t\,dt)$ is the innovations Wiener process.

What conditions must be imposed on the controls will be clarified in due course; however, the theorem holds at least for non-random u_t which is locally bounded (open loop controls), and for sufficiently many feedback controls that we will be able to solve the partial observations counterpart of the linear regulator problem (section 7.5).

In principle it is possible to proceed as we did in the previous section, i.e., by obtaining the Zakai equation through the Kallianpur-Striebel formula. The problem, however, is that the change of measure Λ_t is generally not square-integrable, so we will almost certainly have trouble applying lemma 7.2.7. This can be taken care of by clever localization [Par91] or truncation [Ben92] arguments. We will take an entirely different route, however, which has an elegance of its own: we will exploit the fact that the conditional expectation is the least squares estimate to turn the filtering problem into an optimal control problem (which aims to find an estimator which minimizes the mean square error). The fundamental connection between filtering and control runs very deep (see section 7.7 for references), but is particularly convenient in the Kalman-Bucy case due to the special structure of the linear filtering problem.

Before we embark on this route, let us show that theorem 7.3.1 does indeed follow from the previous section, provided that we are willing to forgo technical precision. The easiest way to do this is to consider the density form of the Zakai equation,

$$\sigma_t(f) \equiv \int f(x)\, q_t(x)\, dx, \quad dq_t(x) = (\mathscr{L}_t^u)^* q_t(x)\, dt + q_t(x)\, (K(t)^{-1} h(t,x))^*\, d\bar{Y}_t.$$

In the linear case, this becomes

$$dq_t(x) = q_t(x)\, (K(t)^{-1} H(t) x)^*\, d\bar{Y}_t + \left[\frac{1}{2} \sum_{i,j=1}^n (C(t)C(t)^*)^{ij}\, \frac{\partial q_t(x)}{\partial x^i \partial x^j} - \sum_{i=1}^n \frac{\partial}{\partial x^i} ((A(t)x + B(t)u_t)^i q_t(x)) \right] dt.$$

You can easily verify, by explicit computation, that

$$q_t(x) = C_t\, \exp\left(-\frac{1}{2} (x - \hat{X}_t)^* \hat{P}_t^{-1} (x - \hat{X}_t) \right)$$

is a solution to the Zakai equation, where C_t is an appropriately chosen non-random function and $\hat{X}_t, \hat{P}_t$ are as given in theorem 7.3.1. Evidently the conditional density of X_t is Gaussian with mean $\hat{X}_t$ and covariance $\hat{P}_t$ (we say that the filtering model is *conditionally Gaussian*), from which theorem 7.3.1 follows directly. To make this approach precise, however, formidable technical problems need to be overcome—does the Zakai equation hold when b and h are unbounded, under what conditions does the Zakai PDE hold, and when are the solutions to these equations unique? (Without the latter, we may have found a solution to the Zakai equation that does not coincide with the optimal filter!) These questions can all be resolved [Ben92, Par91], but, as already announced, we will take a different approach to obtain the result.

Throughout this section we will shamelessly exploit the following lemma.

Lemma 7.3.2. *Denote by $\Phi_{s,t}$ the unique (non-random) matrix that solves*

$$\frac{d}{dt}\,\Phi_{s,t} = A(t)\,\Phi_{s,t} \quad (t > s), \qquad \Phi_{s,s} = I,$$

where I is the identity matrix as usual. Then we can write

$$X_t = \Phi_{0,t}X_0 + \int_0^t \Phi_{s,t}B(s)u_s\,ds + \int_0^t \Phi_{s,t}C(s)\,dW_s.$$

Proof. It is elementary that $\Phi_{s,t} = \Phi_{0,t}(\Phi_{0,s})^{-1}$. Hence the claim is that we can write

$$X_t = \Phi_{0,t}\left[X_0 + \int_0^t (\Phi_{0,s})^{-1}B(s)\,u_s\,ds + \int_0^t (\Phi_{0,s})^{-1}C(s)\,dW_s\right].$$

But this is easily verified by Itô's rule, so we are done. □

The uncontrolled case

We will begin by considering the case where there is no control, i.e., *we assume until further notice that $u_t = 0$*. As it turns out, the linear filtering problem has special structure that will allow us to easily reinsert the controls at the end of the day, so it is convenient not to bother with them in the beginning. In the absence of control, the linear filtering problem has a very special property: (X_t, Y_t) is a *Gaussian process*. This has an important consequence, which will simplify matters considerably.

Lemma 7.3.3. *If $X_n \to X$ in $\mathcal{L}^2$, and X_n is a Gaussian random variable for every n, then X is a Gaussian random variable.*

Proof. Note that $\mathbb{E}(e^{ik^*X_n}) = e^{ik^*\mu_n - k^*P_nk/2}$, where μ_n and P_n are the mean and covariance of X_n. As $X_n \to X$ in $\mathcal{L}^2$, we find that $\mu_n \to \mu$ and $P_n \to P$, where μ and P are the mean and covariance of X (which are finite as $X \in \mathcal{L}^2$). But $\mathbb{E}(e^{ik^*X_n}) \to \mathbb{E}(e^{ik^*X})$ as e^{ik^*x} is bounded and continuous, so $\mathbb{E}(e^{ik^*X}) = e^{ik^*\mu - k^*Pk/2}$. Hence X is Gaussian. □

Lemma 7.3.4. *The finite dimensional distributions of (X_t, Y_t) are Gaussian.*

Proof. Obvious from lemma 7.3.2 and the previous lemma (where $u_t = 0$). □

Why does this help? Recall that we want to compute $\mathbb{E}(X_t|\mathcal{F}_t^Y)$; in general, this could be an arbitrarily complicated measurable functional of the observation sample paths $\{Y_s : s \le t\}$. However, it is a *very special* consequence of the Gaussian property of (X_t, Y_t) that $\mathbb{E}(X_t|\mathcal{F}_t^Y)$ must be a *linear* functional of $\{Y_s : s \le t\}$ (in a sense to be made precise). This will make our life much simpler, as we can easily parametrize all linear functionals of $\{Y_s : s \le t\}$; it then suffices, by the least squares property of the conditional expectation (proposition 2.3.3), to search for the linear functional $\mathcal{L}$ that minimizes the mean square error $\hat{X}_t^i = \operatorname{argmin}_{\mathcal{L}} \mathbb{E}((X_t^i - \mathcal{L}(Y_{[0,t]}))^2)$.

Lemma 7.3.5. *There exists a non-random $\mathbb{R}^{n\times p}$-valued function $G(t,s)$ such that*

$$\mathbb{E}(X_t|\mathcal{F}_t^Y) = \mathbb{E}(X_t) + \int_0^t G(t,s)H(s)(X_s - \mathbb{E}(X_s))\,ds + \int_0^t G(t,s)K(s)\,dB_s,$$

where $\int_0^t \|G(t,s)\|^2 ds < \infty$. Thus $\mathbb{E}(X_t|\mathcal{F}_t^Y)$ is a linear functional of $\{Y_s : s \le t\}$.

Proof. Define the processes

$$\tilde{X}_t = X_t - \mathbb{E}(X_t), \qquad \tilde{Y}_t = \int_0^t H(s)\tilde{X}_s\,dt + \int_0^t K(s)\,dB_s.$$

Clearly $\mathcal{F}_t^Y = \mathcal{F}_t^{\tilde{Y}}$, $(\tilde{X}_t, \tilde{Y}_t)$ is a Gaussian process, and we wish to prove that

$$\mathbb{E}(\tilde{X}_t|\mathcal{F}_t^{\tilde{Y}}) = \int_0^t G(t,s)\,d\tilde{Y}_s = \int_0^t G(t,s)H(s)\tilde{X}_s\,dt + \int_0^t G(t,s)K(s)\,dB_s.$$

Let us first consider a simpler problem which only depends on a finite number of random variables. To this end, introduce the σ-algebra $\mathcal{G}_\ell = \sigma\{Y_{k2^{-\ell}t} - Y_{(k-1)2^{-\ell}t} : k = 1, \ldots, 2^\ell\}$, and note that $\mathcal{F}_t^Y = \sigma\{\mathcal{G}_\ell : \ell = 1, 2, \ldots\}$ (as Y_t has continuous sample paths, so only depends on its values in a dense set of times). Define also the $p2^\ell$-dimensional random variable

$$\tilde{Y}^\ell = (\tilde{Y}_{2^{-\ell}t}, \tilde{Y}_{2\cdot 2^{-\ell}t} - \tilde{Y}_{2^{-\ell}t}, \ldots, \tilde{Y}_t - \tilde{Y}_{(1-2^{-\ell})t}),$$

so that $\mathbb{E}(\tilde{X}_t|\mathcal{G}_\ell) = \mathbb{E}(\tilde{X}_t|\tilde{Y}^\ell)$. But $(\tilde{X}_t, \tilde{Y}^\ell)$ is a $(p2^\ell + n)$-dimensional Gaussian random variable, and in particular possesses a joint (Gaussian) density with respect to the Lebesgue measure. It is well known how to condition multivariate Gaussians, so we will not repeat the computation (it is simply a matter of applying example 7.1.1 to the Gaussian density, and performing explicit integrations); the result is as follows: if we denote by Σ_{XX}, Σ_{YY} the covariance matrices of $\tilde{X}_t$ and $\tilde{Y}^\ell$, and by Σ_{XY} the covariance between $\tilde{X}_t$ and $\tilde{Y}^\ell$, then $\mathbb{E}(\tilde{X}_t|\tilde{Y}^\ell) = \mathbb{E}(\tilde{X}_t) + \Sigma_{XY}(\Sigma_{YY})^{-1}(\tilde{Y}^\ell - \mathbb{E}(\tilde{Y}^\ell))$ (if Σ_{YY} is singular, take the pseudoinverse instead). But for us $\mathbb{E}(\tilde{X}_t) = \mathbb{E}(\tilde{Y}^\ell) = 0$, so we conclude that $\mathbb{E}(\tilde{X}_t|\tilde{Y}^\ell) = \Sigma_{XY}(\Sigma_{YY})^{-1}\tilde{Y}^\ell$.

Evidently $\mathbb{E}(\tilde{X}_t|\tilde{Y}^\ell)$ can be written as a linear combination of the increments of $\tilde{Y}^\ell$ with deterministic coefficients. In particular, we can thus write

$$\mathbb{E}(\tilde{X}_t|\mathcal{G}_\ell) = \int_0^t G^\ell(t,s)\,d\tilde{Y}_s = \int_0^t G^\ell(t,s)H(s)\tilde{X}_s\,dt + \int_0^t G^\ell(t,s)K(s)\,dB_s,$$

where $s \mapsto G^\ell(t,s)$ is a non-random simple function which is constant on the intervals $s \in [(k-1)2^{-\ell}t, k2^{-\ell}t[$. To proceed, we would like to take the limit as $\ell \to \infty$. But note that $\mathbb{E}(\tilde{X}_t|\mathcal{G}_\ell) \to \mathbb{E}(\tilde{X}_t|\mathcal{F}_t^Y)$ in $\mathcal{L}^2$ by Lévy's upward theorem (lemma 4.6.4). Hence the remainder is essentially obvious (see [LS01a, lemma 10.1] for more elaborate reasoning). □

We can now proceed to solve the filtering problem. Our task is clear: out of all linear functionals of the form defined in the previous lemma, we seek the one that minimizes the mean square error. We will turn this problem into an optimal control problem, for which $G(t,s)$ in lemma 7.3.5 is precisely the optimal control.

Theorem 7.3.6 (Kalman-Bucy, no control). *Theorem 7.3.1 holds for $u_t = 0$.*

Proof. Let us fix the terminal time T. For any (non-random) function $G : [0,T] \to \mathbb{R}^{n\times p}$ with $\int_0^T \|G(t)\|^2 dt < \infty$, we define the $\mathcal{F}_t^Y$-adapted process

$$L_t^G = \mathbb{E}(X_t) + \int_0^t G(s)H(s)(X_s - \mathbb{E}(X_s))\,ds + \int_0^t G(s)K(s)\,dB_s,$$

We would like to find such a G that minimizes the cost $J_v[G] = \mathbb{E}((v^*(X_T - L_T^G))^2)$ for every vector $v \in \mathbb{R}^n$. By proposition 2.3.3 and lemma 7.3.5, $v^*L_T^{G^*} = \mathbb{E}(v^*X_T|\mathcal{F}_T^Y) = v^*\hat{X}_T$ and $J_v[G_*] = v^*\hat{P}_T v$ for every $v \in \mathbb{R}^n$, where G_* is the function that minimizes $J_v[G]$.

Somewhat surprisingly, this is a linear regulator problem in disguise; we have to do a little work to make this explicit. The idea is to obtain an equation for X_T, in almost the same way as we did in lemma 7.3.2, such that the resulting expression for $X_T - L_T^G$ contains only stochastic integrals and no time integrals. Then, using the Itô isometry, we find an expression for $J_v[G]$ that looks like the quadratic cost in the linear regulator problem. To this end, define

$$\frac{d}{ds}\Psi_{s,t}^G + \Psi_{s,t}^G A(t) - G(t)H(t) = 0 \quad (s < t), \qquad \Psi_{t,t}^G = I.$$

Applying Itô's rule to $\Psi_{t,T}^G(X_t - \mathbb{E}(X_t))$ gives

$$X_T - \mathbb{E}(X_T) = \Psi_{0,T}^G(X_0 - \mathbb{E}(X_0)) + \int_0^T G(s)H(s)(X_s - \mathbb{E}(X_s))\,ds + \int_0^T \Psi_{s,T}^G C(s)\,dW_s.$$

This gives the following expression fo $X_T - L_T^G$:

$$X_T - L_T^G = \Psi_{0,T}^G(X_0 - \mathbb{E}(X_0)) + \int_0^T \Psi_{s,T}^G C(s)\,dW_s - \int_0^T G(s)K(s)\,dB_s.$$

This expression has no time integrals, as desired. We now easily compute

$$J_v[G] = \int_0^T \left\{ \|C(s)^*(\Psi_{s,T}^G)^* v\|^2 + \|K(s)^* G(s)^* v\|^2 \right\} ds + v^* \Psi_{0,T}^G \hat{P}_0 (\Psi_{0,T}^G)^* v.$$

Now define $\alpha(t) = G(T-t)^* v$, $\xi_t^\alpha = (\Psi_{T-t,T}^G)^* v$, so that we can write

$$\frac{d}{dt}\xi_t^\alpha = A(T-t)^* \xi_t^\alpha - H(T-t)^* \alpha(t), \qquad \xi_0^\alpha = v,$$

and we obtain the cost $J_v[G] = J[\alpha]$ with

$$J_v[G] = J[\alpha] = \int_0^T \{(\xi_s^\alpha)^* C(T-s)C(T-s)^* \xi_s^\alpha + \alpha(s)^* K(T-s)K(T-s)^* \alpha(s)\}\,ds + (\xi_T^\alpha)^* \hat{P}_0 \xi_T^\alpha.$$

But this is precisely a linear regulator problem for the controlled (non-random) differential equation ξ_t^α with the cost $J[\alpha]$. The conclusion of the theorem follows easily (fill in the remaining steps!) by invoking the solution of the linear regulator problem (theorem 6.5.1).

It remains to verify that the innovations process $\bar{B}_t$ is a Wiener process, as claimed; this follows immediately, however, from proposition 7.2.9, without any changes in the proof. □

The controlled case

Now that we have obtained the Kalman-Bucy filter without control, it remains to consider the controlled case. Once again, the linear structure of the problem simplifies matters considerably; it allows us to infer the solution of the controlled filtering problem from its uncontrolled counterpart, which we have alreay solved! It should be clear that we could never hope for such a result in the general case, but as the linear filtering problem has such nice structure we will be happy to shamelessly exploit it.

What is the idea? Recall that, by lemma 7.3.2

$$X_t^u = \Phi_{0,t}X_0 + \int_0^t \Phi_{s,t}B(s)u_s\,ds + \int_0^t \Phi_{s,t}C(s)\,dW_s,$$

where we have attached the label ^{u} to the signal process to denote its solution with the control strategy u in operation. But then evidently

$$X_t^u = X_t^0 + \int_0^t \Phi_{s,t} B(s) u_s \, ds,$$

and in particular the second term is $\mathcal{F}_t^{Y,u}$-adapted (as u_s was assumed to depend only on the observations), where we have denoted the observations under the strategy u by

$$Y_t^u = \int_0^t H(s) X_s^u \, ds + \int_0^t K(s) \, dB_s, \qquad \mathcal{F}_t^{Y,u} = \sigma\{Y_s^u : s \le t\}.$$

Hence we obtain

$$\hat{X}_t^u = \mathbb{E}(X_t^u | \mathcal{F}_t^{Y,u}) = \mathbb{E}(X_t^0 | \mathcal{F}_t^{Y,u}) + \int_0^t \Phi_{s,t} B(s) u_s \, ds.$$

If only $\mathcal{F}_t^{Y,u} = \mathcal{F}_t^{Y,0}$, we would easily obtain an equation for $\hat{X}_t^u$: after all, then

$$\hat{X}_t^u = \hat{X}_t^0 + \int_0^t \Phi_{s,t} B(s) u_s \, ds,$$

and as we have already found the equation for $\hat{X}_t^0$ we immediately obtain the appropriate equation for $\hat{X}_t^u$ using Itô's rule. The statement $\mathcal{F}_t^{Y,u} = \mathcal{F}_t^{Y,0}$ is not at all obvious, however. The approach which we will take is simply to restrict consideration only to those control strategies for which $\mathcal{F}_t^{Y,u} = \mathcal{F}_t^{Y,0}$ is satisfied; we will subsequently show that this is indeed the case for a large class of interesting controls.

Remark 7.3.7. We had no such requirement in the bounded case, where we used the Kallianpur-Striebel formula to obtain the filter. Indeed this requirement is also superfluous here: it can be shown that under a straightforward integrability condition (of purely technical nature), $\mathbb{E}(X_t^0 | \mathcal{F}_t^{Y,u}) = \mathbb{E}(X_t^0 | \mathcal{F}_t^{Y,0})$ always holds regardless of whether $\mathcal{F}_t^{Y,u} = \mathcal{F}_t^{Y,0}$ [Ben92, section 2.4]. It is perhaps not surprising that the proof of this fact hinges crucially on the Kallianpur-Striebel formula! We will not need this level of generality, however, as it turns out that $\mathcal{F}_t^{Y,u} = \mathcal{F}_t^{Y,0}$ for a sufficiently large class of controls; we will thus be content to stick to this simpler approach.

The following result is now basically trivial.

Theorem 7.3.8 (Kalman-Bucy with control). *Suppose that* $u_\cdot \in \bigcap_{T<\infty} \mathcal{L}^1(\mu_T \times \mathbb{P})$ *and that* $\mathcal{F}_t^{Y,u} = \mathcal{F}_t^{Y,0}$ *for all* $t < \infty$. *Then theorem 7.3.1 holds.*

Proof. The integrability condition ensures that $X_t^u - X_t^0$ is in $\mathcal{L}^1$ (so that the conditional expectation is well defined). The discussion above gives immediately the equation for $\hat{X}_t^u$, which depends on $\hat{P}_t$. We claim that $\hat{P}_t = \mathbb{E}((X_t^u - \hat{X}_t^u)(X_t^u - \hat{X}_t^u)^*)$; but this follows immediately from $X_t^u - \hat{X}_t^u = X_t^0 - \hat{X}_t^0$. It remains to show that the innovation $\bar{B}_t^u$ is still a Wiener process. But as $X_t^u - \hat{X}_t^u = X_t^0 - \hat{X}_t^0$, we find that $\bar{B}_t^u = \bar{B}_t^0$, which we have already established to be a Wiener process. Hence the proof is complete. □

This result is not very useful by itself; it is not clear that there even exist controls that satisfy the conditions! (Of course, non-random controls are easily seen to work, but these are not very interesting.) We will conclude this section by exhibiting two classes of controls that satisfy the conditions of theorem 7.3.8.

The first, and the most important class in the following, consists of those controls whose value at time t is a Lipschitz function of the Kalman-Bucy filter at time t. This class of *separated controls* is particularly simple to implement: we can update the Kalman-Bucy filter numerically, e.g., using the Euler-Maruyama method, and at each time we simply feed back a function of the latest estimate. In particular, the complexity of feedback strategies that depend on the entire obsevation path in an arbitrary way is avoided. As it turns out, controls of this form are also optimal for the type of optimal control problems in which we are interested (see section 7.5).

Proposition 7.3.9. *Let $u_t = \alpha(t, \hat{X}_t)$, where $\alpha : [0, \infty[\times \mathbb{R}^n \to \mathbb{R}^k$ is Lipschitz and*

$$d\hat{X}_t = A(t)\hat{X}_t\,dt + B(t)\alpha(t,\hat{X}_t)\,dt + \hat{P}_t(K(t)^{-1}H(t))^*K(t)^{-1}(dY_t^u - H(t)\hat{X}_t\,dt)$$

with the initial condition $\hat{X}_0 = \hat{X}_0$. Then $(X_t^u, Y_t^u, \hat{X}_t)$ has a unique solution, u_t satisfies the conditions of theorem 7.3.8, and $\hat{X}_t = \hat{X}_t^u$.

Proof. Set $F(t) = K(t)^{-1}H(t)$. To see that $(X_t^u, Y_t^u, \hat{X}_t)$ has a unique solution, write

$$\begin{aligned}
dX_t^u &= (A(t)X_t^u + B(t)\alpha(t,\hat{X}_t))\,dt + C(t)\,dW_t,\\
dY_t^u &= H(t)X_t^u\,dt + K(t)\,dB_t,\\
d\hat{X}_t &= (A(t)\hat{X}_t + B(t)\alpha(t,\hat{X}_t))\,dt + \hat{P}_tF(t)^*(dB_t + F(t)(X_t^u - \hat{X}_t)\,dt),
\end{aligned}$$

and note that this SDE has Lipschitz coefficients. Existence and uniqueness follows from theorem 5.1.3, as well as the fact that $u_t = \alpha(t, \hat{X}_t)$ is in $\mathcal{L}^2(\mu_T \times \mathbb{P})$ for all T.

To proceed, consider the unique solution to the equation

$$dX_t' = (A(t)X_t' + B(t)\alpha(t,X_t'))\,dt + \hat{P}_tF(t)^*d\bar{B}_t^0, \qquad X_0' = \hat{X}_0,$$

which is $\mathcal{F}_t^{Y,0}$-adapted as $\bar{B}_t^0$ is an $\mathcal{F}_t^{Y,0}$-Wiener process and the coefficients are Lipschitz. Consider the $\mathcal{F}_t^{Y,0}$-adapted control $u_t' = \alpha(t, X_t')$. It is easily seen that

$$\mathbb{E}(X_t^{u'}|\mathcal{F}_t^{Y,0}) = \hat{X}_t^0 + \int_0^t \Psi_{s,t}B(s)u_s'\,ds.$$

Using Itô's rule, we find that $\mathbb{E}(X_t^{u'}|\mathcal{F}_t^{Y,0})$ satisfies the same equation as X_t', so apparently $X_t' = \mathbb{E}(X_t^{u'}|\mathcal{F}_t^{Y,0})$ by the uniqueness of the solution. On the other hand, note that

$$d\bar{B}_t^0 = K(t)^{-1}dY_t^0 - F(t)\hat{X}_t^0\,dt = K(t)^{-1}dY_t^{u'} - F(t)X_t'\,ds,$$

so we can write

$$dX_t' = (A(t)X_t' + B(t)\alpha(t,X_t'))\,dt + \hat{P}_tF(t)^*(K(t)^{-1}dY_t^{u'} - F(t)X_t'\,ds).$$

Thus X_t' is $\mathcal{F}_t^{Y,u'}$-adapted (e.g., note that $X_t' - \int_0^t \hat{P}_sF(s)^*K(s)^{-1}dY_s^{u'}$ satisfies an ODE which has a unique solution), so $u_t' = \alpha(t, X_t')$ is both $\mathcal{F}_t^{Y,0}$- and $\mathcal{F}_t^{Y,u'}$-adapted. But note that

$$Y_t^{u'} = Y_t^0 + \int_0^t H(s)\int_0^s \Phi_{r,t}B(r)u_r'\,dr\,ds.$$

Thus $\mathcal{F}_t^{Y,0} \subset \mathcal{F}_t^{Y,u'}$, as Y_t^0 is a functional of $Y_t^{u'}$ and u'_t, both of which are $\mathcal{F}_t^{Y,u'}$-adapted. Conversely $\mathcal{F}_t^{Y,u'} \subset \mathcal{F}_t^{Y,0}$, as $Y_t^{u'}$ is a functional of Y_t^0 and u'_t, both of which are $\mathcal{F}_t^{Y,0}$-adapted. It remains to note that $(X_t^{u'}, Y_t^{u'}, X'_t)$ satisfies the same SDE as $(X_t^u, Y_t^u, \hat{X}_t)$, so $u'_t = u_t$, etc., by uniqueness, and in particular $\tilde{X}_t = X'_t = \mathbb{E}(X_t^{u'}|\mathcal{F}_t^{Y,0}) = \hat{X}_t^{u'} = \hat{X}_t^u$. □

The second class of controls that are guaranteed to satisfy the conditions of theorem 7.3.8 consists of those strategies which are "nice" but otherwise arbitrary functionals of the observation history. The following result is not difficult to obtain, but as we will not need it we refer to [FR75, lemma VI.11.3] for the proof. Let us restrict to a finite interval $[0, T]$ for notational simplicity (the extension to $[0, \infty[$ is trivial).

Proposition 7.3.10. *Let $\alpha : [0,T] \times C([0,T];\mathbb{R}^p) \to \mathbb{R}^k$ be a (Borel-)measurable function which satisfies the following condtions:*

1. *If $y_s = y'_s$ for all $s \le t$, then $\alpha(t,y) = \alpha(t,y')$; in other words, $\alpha(t,y)$ only depends on $\{y_s : s \le t\}$ (for fixed t).*
2. *$\|\alpha(t,y) - \alpha(t,y')\| \le K \max_{s\in[0,T]} \|y_s - y'_s\|$ for some $K < \infty$ and all $t \in [0,T]$ and $y, y' \in C([0,T];\mathbb{R}^p)$; i.e., the function α is uniformly Lipschitz.*
3. *$\|\alpha(t,0)\|$ is bounded on $t \in [0,T]$.*

Define the control $u_t = \alpha(t, Y_\cdot)$. Then the equation for (X_t, Y_t) admits a unique solution which satisfies the conditions of theorem 7.3.8.

7.4 The Shiryaev-Wonham filter

Beside the Kalman filter, there is another important class of finite-dimensionally computable filters. Unlike in the previous sections, the signal process in these filtering models is not defined as the solution of a stochastic differential equation. Instead, one considers signal processes which take a finite number of values (and hence have piecewise constant sample paths)—in particular, finite-dimensional filters arise in the case that the signal process is any *finite state continuous time Markov process*, and the resulting filters are called *Wonham filters* [Won65, LS01a]. You can imagine why this simplifies matters: if X_t only takes a finite number of values $x_1, \ldots, x_n$ at every time t, then the knowledge of $\mathbb{P}(X_t = x_i|\mathcal{F}_t^Y)$, $i = 1, \ldots, n$ is sufficient to compute any filtered estimate $\pi_t(f)$. Hence the Wonham filter is a finite-dimensional SDE, driven by the observations, which propagates the n-dimensional vector $\pi_t^i = \mathbb{P}(X_t = x_i|\mathcal{F}_t^Y)$. In a discrete time setting, we have encountered exactly the same idea in example 7.2.10.

Developing the general theory for such filters is not at all difficult, but requires some knowledge of continuous time Markov chains. Rather than going in this direction, we will discuss a particularly straightforward special case which dates back to the early work of Shiryaev [Shi63, Shi73]. We will encounter this filter again in the next chapter, where it will be combined with optimal stopping theory to develop some interesting applications in statistics and in finance.

We consider the following model. We work on the space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, \mathbb{P})$, on which is defined an $\mathcal{F}_t$-Wiener process B_t and an $\mathcal{F}_0$-measurable random variable τ with values in $[0, \infty]$. We will assume that τ is distributed as follows:

$$\mathbb{P}(\tau = 0) = p_0, \qquad \mathbb{P}(\tau = \infty) = p_\infty, \qquad \mathbb{P}(\tau \le t | 0 < \tau < \infty) = \int_0^t p_\tau(s)\, ds,$$

where $0 \le p_0, p_\infty \le 1$, $p_0 + p_\infty \le 1$, and $p_\tau(s)$ (the density of the continuous part of τ) is a nonnegative function such that $\int_0^\infty p_\tau(s) ds = 1$. You should interpret τ as the random time at which a sudden change occurs in our system, e.g., a system failure of some kind. Then p_0 is the probability that the change occured before we start observing, p_∞ is the probability that the change never occurs, and the probability that the change occurs in the interval $]0, t]$ is $(1 - p_0 - p_\infty)\mathbb{P}(\tau \le t | 0 < \tau < \infty)$.

Unfortunately, we cannot see directly when the change occurs. Instead, we only have access to noisy observations of the form $y_t = \gamma\, I_{\tau \le t} + \sigma\, \xi_t$, where ξ_t is white noise; as usual, we will work with the integrated observations

$$Y_t = \gamma \int_0^t I_{\tau \le s}\, ds + \sigma\, B_t.$$

The goal of the filtering problem is to estimate whether the change has occured by the current time, given the observations up to the current time; in other words, we seek to compute $\pi_t = \mathbb{P}(\tau \le t | \mathcal{F}_t^Y) = \mathbb{E}(I_{\tau \le t} | \mathcal{F}_t^Y)$, where $\mathcal{F}_t^Y = \sigma\{Y_s : s \le t\}$.

We have all the tools to solve this problem; in fact, compared to some of the more technical problems which we have encountered in the previous sections, this is a piece of cake! All that is needed is the Bayes formula and some simple manipulations.

Proposition 7.4.1 (Shiryaev-Wonham filter). *$\pi_t = \mathbb{P}(\tau \le t | \mathcal{F}_t^Y)$ satisfies*

$$d\pi_t = \frac{\gamma}{\sigma}\, \pi_t(1 - \pi_t)\, d\bar{B}_t + \frac{(1 - p_0 - p_\infty)\, p_\tau(t)}{(1 - p_0 - p_\infty) \int_t^\infty p_\tau(s)\, ds + p_\infty} (1 - \pi_t)\, dt, \quad \pi_0 = p_0,$$

where the innovations process $d\bar{B}_t = \sigma^{-1}(dY_t - \gamma \pi_t\, dt)$ is an $\mathcal{F}_t^Y$-Wiener process.

Proof. Consider the following change of measure:

$$\frac{d\mathbb{Q}_T}{d\mathbb{P}} = \exp\left(-\frac{\gamma}{\sigma} \int_0^T I_{\tau \le s}\, dB_s - \frac{\gamma^2}{2\sigma^2} \int_0^T I_{\tau \le s}\, ds \right).$$

Clearly Novikov's condition is satisfied, so by Girsanov's theorem $\sigma^{-1} Y_t$ is an $\mathcal{F}_t$-Wiener process under $\mathbb{Q}_T$ on the interval $[0, T]$. In particular τ and Y_t are independent, and it is easily verified that τ has the same law under $\mathbb{Q}_T$ and under $\mathbb{P}$. Now note that we obtain

$$\frac{d\mathbb{P}}{d\mathbb{Q}_t} = \exp\left(\frac{\gamma}{\sigma^2}(Y_t - Y_{t \wedge \tau}) - \frac{\gamma^2}{2\sigma^2}(t - \tau)^+ \right)$$

after some simple manipulations. Thus, by the Bayes formula and independence of τ and Y_t,

$$\pi_t = \mathbb{P}(\tau \le t | \mathcal{F}_t^Y) = \frac{\int_{[0,\infty]} I_{s \le t} \exp(\frac{\gamma}{\sigma^2}(Y_t - Y_{t \wedge s}) - \frac{\gamma^2}{2\sigma^2}(t - s)^+)\, \mu_\tau(ds)}{\int_{[0,\infty]} \exp(\frac{\gamma}{\sigma^2}(Y_t - Y_{t \wedge s}) - \frac{\gamma^2}{2\sigma^2}(t - s)^+)\, \mu_\tau(ds)},$$

where μ_τ is the law of τ. Let us now evaluate the numerator and denominator explicitly. For the numerator, we find the explicit expression

$$\Sigma_t = \int_{[0,\infty]} I_{s\le t}\, \exp(\tfrac{\gamma}{\sigma^2}(Y_t - Y_{t\wedge s}) - \tfrac{\gamma^2}{2\sigma^2}(t-s)^+)\, \mu_\tau(ds) = \\ p_0\, e^{\frac{\gamma}{\sigma^2}Y_t - \frac{\gamma^2}{2\sigma^2}t} + (1-p_0-p_\infty)\int_0^t p_\tau(s)\, e^{\frac{\gamma}{\sigma^2}(Y_t-Y_s)-\frac{\gamma^2}{2\sigma^2}(t-s)}\, ds.$$

On the other hand, it is easy to see that for the denominator

$$\int_{[0,\infty]} \exp(\tfrac{\gamma}{\sigma^2}(Y_t - Y_{t\wedge s}) - \tfrac{\gamma^2}{2\sigma^2}(t-s)^+)\, \mu_\tau(ds) = \Sigma_t + (1-p_0-p_\infty)\int_t^\infty p_\tau(s)\, ds + p_\infty.$$

It remains to apply Itô's rule. First, applying Itô's rule to Σ_t, we obtain the counterpart of the Zakai equation in the current context:

$$d\Sigma_t = \frac{\gamma}{\sigma^2}\,\Sigma_t\, dY_t + (1-p_0-p_\infty)\, p_\tau(t)\, dt, \qquad \Sigma_0 = p_0.$$

Another application of Itô's rule gives

$$d\pi_t = \frac{\gamma}{\sigma}\,\pi_t(1-\pi_t)\,\sigma^{-1}(dY_t - \gamma\pi_t\, dt) + \frac{(1-p_0-p_\infty)\, p_\tau(t)}{(1-p_0-p_\infty)\int_t^\infty p_\tau(s)\, ds + p_\infty}\,(1-\pi_t)\, dt.$$

It remains to note that $d\bar{B}_t = \sigma^{-1}(dY_t - \gamma\pi_t\, dt)$ is a Wiener process, which follows exactly as in proposition 7.2.9 without any change to the proof. □

Remark 7.4.2. The uniqueness of the solution of the Shiryaev-Wonham equation is not entirely obvious, as its coefficients do not satisfy the global Lipschitz condition. However, they do satisfy the local Lipschitz condition, so have unique solutions until an explosion time ζ (see section 5.6). On the other hand, we know that there exists at least one solution $\pi_t = \mathbb{P}(\tau \le t|\mathcal{F}_t^Y)$ which, by construction, remains in the interval $[0,1]$ forever. Hence it must be the case that $\zeta = \infty$ a.s.[2] This is important, as it means that we can actually use the Shiryaev-Wonham equation to compute the filtered estimate π_t (this would not be obvious if the solution were not unique).

An important special case is the setting in which τ is exponentially distributed, i.e., $p_\tau(t) = \lambda\, e^{-\lambda t}$ for some $\lambda > 0$ and $p_\infty = 0$. The particular relevance of this choice is that then $I_{\tau\le t}$ becomes a *time-homogeneous* Markov process, which manifests itself by the fact that the Shiryaev-Wonham equation becomes time-homogeneous:

$$d\pi_t = \frac{\gamma}{\sigma}\,\pi_t(1-\pi_t)\,\sigma^{-1}(dY_t - \gamma\pi_t\, dt) + \lambda\,(1-\pi_t)\, dt, \qquad \pi_0 = p_0.$$

Models with exponential waiting times are common in applications; they correspond to the situation where the change is equally likely to occur in every time interval of fixed length (as $\mathbb{P}(\tau \in]t, t+\Delta] \,|\, \tau > t) = 1 - e^{-\lambda\Delta}$). This setting will be particularly convenient in combination with optimal stopping theory in the next chapter.

[2] This is only true, of course, provided that we start with $\pi_0 \in [0,1]$—the estimate π_t must be a probability! In the meaningless scenario, at least from a filtering perspective, where we try to solve the Shiryaev-Wonham equation for $\pi_0 \notin [0,1]$, the solutions may indeed explode in finite time.

7.5 The separation principle and LQG control

Let us now move on from filtering, and investigate systems with noisy observations in the setting of optimal control. Consider again the system-observation pair

$$\begin{aligned} dX_t^u &= b(t, X_t^u, u_t)\,dt + \sigma(t, X_t^u, u_t)\,dW_t, \\ dY_t^u &= h(t, X_t^u, u_t)\,dt + K(t)\,dB_t. \end{aligned}$$

We would like to design a strategy u_t to achieve a certain purpose; consider, for example, a cost functional that is similar to the finite horizon cost in the previous chapter:

$$J[u] = \mathbb{E}\left[\int_0^T \{v(X_t^u) + w(u_t)\}\,dt + z(X_T^u)\right].$$

(The specific type of running cost is considered for simplicity only, and is certainly not essential.) Our goal is, as usual, to find a control strategy u^* that minimizes the cost $J[u]$. However, as opposed to the previous chapter, there is now a new ingredient in the problem: we can only observe the noisy sensor data Y_t^u, so that the control signal u_t can only be $\mathcal{F}_t^{Y,u}$-adapted (where $\mathcal{F}_t^{Y,u} = \sigma\{Y_s^u : s \le t\}$). The theory of the previous chapter cannot account for this; the only constraint which we are able to impose on the control signal within that framework is the specification of the control set $\mathbb{U}$, and the constraint that u_t is $\mathcal{F}_t^{Y,u}$-adapted is certainly not of this form. Indeed, if we apply the Bellman equation, we always automatically obtain a Markov control which is a function of X_t^u and is thus not adapted to the observations.

The trick to circumvent this problem is to express the cost in terms of quantities that depend only on the observations; if we can then find a feedback control which is a function of those quantities, then that control is automatically $\mathcal{F}_t^{Y,u}$-adapted! It is not difficult to express the cost in terms of observation-dependent quantities; indeed, using lemma 7.2.7 and the tower property of the conditional expectation,

$$J[u] = \mathbb{E}\left[\int_0^T \{\pi_t^u(v) + w(u_t)\}\,dt + \pi_T^u(z)\right]$$

(provided we assume that u_t is $\mathcal{F}_t^{Y,u}$-adapted and that we have sufficient integrability to apply lemma 7.2.7), where $\pi_t^u(f) = \mathbb{E}(f(X_t^u)|\mathcal{F}_t^{Y,u})$. But we can now interpret this cost as defining a new control problem, where the system X_t^u is replaced by the filter $\pi_t^u(\cdot)$, and, from the point of view of the filter, we end up with a completely observed optimal control problem. If we can solve a Bellman equation for such a problem, then the optimal control at time t will simply be some function of the filter at time t. Note that this is *not* a Markov control from the point of view of the physical system X_t^u, but this *is* a Markov control from the point of view of the filter. The idea to express the control problem in terms of the filter is often referred to as the *separation principle*, and a strategy which is a function of the filter is called a *separated control*.

Remark 7.5.1. You might worry that we cannot consider the filter by itself as an autonomous system to be controlled, as the filter is driven by the observations Y_t^u

obtained from the physical system rather than by a Wiener proces as in our usual control system models. But recall that the filter can also be expressed in terms of the innovations process: from this point of view, the filter looks just like an autonomous equation, and can be considered as a stochastic differential equation quite separately from the underlying model from which it was obtained.

Unfortunately, the separation principle does not in general lead to results that are useful in practice. We have already seen that in most cases, the filter cannot be computed in a finite-dimensional fashion. At the very best, then, the separation principle leads to an optimal control problem for a stochastic PDE. Even if the *formidable* technical problems along the way can all be resolved (to see that tremendous difficulties will be encountered requires little imagination), this is still of essentially no practical use; after all, an implementation of the controller would require us both to propagate a stochastic PDE in real time, and to evaluate a highly complicated function (the control function) on an infinite-dimensional space! The former is routinely done in a variety of applications, but the latter effectively deals the death blow to applications of *optimal* control theory in the partially observed setting.[3]

On the other hand, in those cases where the filtering problem admits a finite-dimensional solution, the separation principle becomes a powerful tool for control design. In the remainder of this section we will develop one of the most important examples: the partially observed counterpart of the linear regulator problem. We will encounter more applications of the separation principle in the next chapter.

We consider the system-observation model

$$\begin{aligned} dX_t^u &= A(t)X_t^u\,dt + B(t)u_t\,dt + C(t)\,dW_t, \\ dY_t^u &= H(t)X_t^u\,dt + K(t)\,dB_t, \end{aligned}$$

where the various objects in this expression satisfy the same conditions as in section 7.3. Our goal is to find a control strategy u which minmizes the cost functional

$$J[u] = \mathbb{E}\left[\int_0^T \{(X_t^u)^* P(t) X_t^u + (u_t)^* Q(t) u_t\}\,dt + (X_T^u)^* R(X_T^u)\right],$$

where $P(t)$, $Q(t)$ and R satisfy the same conditions as in section 6.5. We will insist, however, that the control strategy u_t is $\mathcal{F}_t^{Y,u}$-adapted, and we seek a strategy that is optimal within the class of such controls that satisfy the conditions of theorem 7.3.8. This is the *LQG (Linear, Quadratic cost, Gaussian) control problem*.

Theorem 7.5.2 (LQG control). *Denote by N_t the solution of the Riccati equation*

$$\frac{dN_t}{dt} = A(t)N_t + N_t A(t)^* - N_t H(t)^* (K(t)K(t)^*)^{-1} H(t) N_t + C(t)C(t)^*,$$

[3] That is not to say that this setting has not been studied; many questions of academic interest, e.g., on the existence of optimal controls, have been investigated extensively. However, I do not know of a single practical application where the separation principle has actually been applied in the infinite-dimensional setting; the optimal control problem is simply too difficult in such cases, and other solutions must be found.

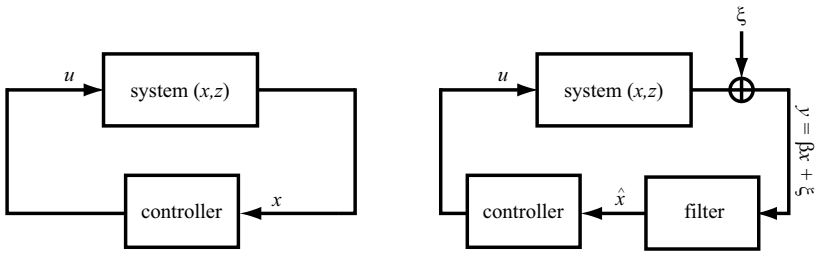

Figure 7.2. Figure 0.3 revisited. The schematic on the left depicts the structure of a completely observed optimal controller, as in the linear regulator problem. The schematic on the right depicts the structure of a separated controller, as in the LQG problem.

where the initial condition N_0 is taken to be the covariance matrix of X_0, and denote by M_t the solution of the time-reversed Riccati equation

$$\frac{d}{dt}M_t + A(t)^*M_t + M_tA(t) - M_tB(t)Q(t)^{-1}B(t)^*M_t + P(t) = 0,$$

with the terminal condition $M_T = R$. Then an optimal feedback control strategy for the LQG control problem is given by $u_t^* = -Q(t)^{-1}B(t)^*M_t\hat{X}_t$, where

$$\begin{aligned} d\hat{X}_t = {} & (A(t) - B(t)Q(t)^{-1}B(t)^*M_t)\hat{X}_t\,dt \\ & + N_t(K(t)^{-1}H(t))^*K(t)^{-1}(dY_t^{u^*} - H(t)\hat{X}_t\,dt), \qquad \hat{X}_0 = \mathbb{E}(X_0), \end{aligned}$$

and $\hat{X}_t = \hat{X}_t^{u^*}$, $N_t = \hat{P}_t$ are the optimal estimate and error covariance for $X_t^{u^*}$.

Proof. As we assume that our controls satisfy the conditions of theorem 7.3.8, the Kalman-Bucy filtering equations are valid. We would thus like to express the cost $J[u]$ in terms of the Kalman-Bucy filter. To this end, note that for any (non-random) matrix G

$$\mathbb{E}((X_t^u)^*GX_t^u) - \mathbb{E}((\hat{X}_t^u)^*G\hat{X}_t^u) = \mathbb{E}((X_t^u - \hat{X}_t^u)^*G(X_t^u - \hat{X}_t^u)) = \mathrm{Tr}[G\hat{P}_t].$$

Thus evidently, the following cost differs from $J[u]$ only by terms that depend on $\mathrm{Tr}[G\hat{P}_t]$ (provided we assume that u_t is a functional of the observations):

$$J'[u] = \mathbb{E}\left[\int_0^T \{(\hat{X}_t^u)^*P(t)\hat{X}_t^u + (u_t)^*Q(t)u_t\}\,dt + (\hat{X}_T^u)^*R(\hat{X}_T^u)\right].$$

But $\mathrm{Tr}[G\hat{P}_t]$ is non-random and does not depend on the control u, so that clearly a strategy u^* which minimizes $J'[u]$ will also minimize $J[u]$. Now note that $\hat{X}_t^u$ satisfies the equation

$$d\hat{X}_t^u = A(t)\hat{X}_t^u\,dt + B(t)u_t\,dt + \hat{P}_t(K(t)^{-1}H(t))^*d\bar{B}_t,$$

where $\bar{B}_t$ is a Wiener process. Hence the equation $\hat{X}_t^u$, together with the cost $J'[u]$, defines a linear regulator problem. By theorem 6.5.1, we find that an optimal control is given by the strategy $u_t^* = -Q(t)^{-1}B(t)^*M_t\hat{X}_t^{u^*}$, and this control is admissible in the current setting by proposition 7.3.9. Moreover, the controlled Kalman-Bucy filter is given precisely by $\hat{X}_t$ in the statement of the theorem. Hence we are done. □

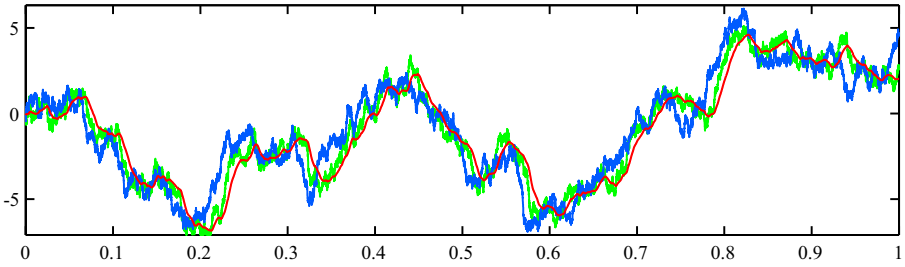

Figure 7.3. Simulation of the model of example 7.5.4 with the optimal control strategy in operation. Shown are the position of the particle x_t (blue), the best estimate of the particle position $\hat{x}_t$ (green), and the position of the microscope focus $-z_t$ (red). For this simulation $T = 1$, $\beta = 100$, $\sigma = 10$, $\gamma = 5$, $\kappa = .5$, $z_0 = 0$, $\mathbb{E}(x_0) = 0$, $\mathrm{var}(x_0) = 2$, $P = 1$, $Q = .5$.

Remark 7.5.3. It is worth pointing out once again the structure of the controls obtained through the separation principle (figure 7.2). In the completely observed case (e.g., the linear regulator), the controller has access to the state of the system, and computes the feedback signal as a memoryless function of the system state. In the partially observed case (e.g., the LQG problem), the noisy observations are first used to compute the best estimate of the system state; the controller then feeds back a memoryless function of this estimate. Evidently the optimal control strategy *separates* into a filtering step and a memoryless controller, hence the name "separation principle".

Example 7.5.4 (Tracking under a microscope IV). Let us return for the last time to our tracking example. In addition to the previous model, we now have observations:

$$\frac{dz_t}{dt} = \beta u_t, \qquad x_t = x_0 + \sigma W_t, \qquad dy_t = \gamma\,(x_t + z_t)\,dt + \kappa\,dB_t.$$

We would like to find a strategy u^* which minimizes the cost

$$J[u] = \mathbb{E}\left[P\int_0^T (x_t + z_t)^2\,dt + Q\int_0^T (u_t)^2\,dt\right],$$

but this time we are only allowing our controls to depend on the noisy observations. As before we will define $e_t = x_t + z_t$, so that we can work with the system equation $de_t = \beta u_t\,dt + \sigma\,dW_t$ (as the observations and cost functional depend only on e_t). But we can now directly apply theorem 7.5.2. We find that $u_t^* = -Q^{-1}\beta\,m_t\,\hat{e}_t$, where

$$d\hat{e}_t = -\frac{\beta^2\,m_t}{Q}\,\hat{e}_t\,dt + \frac{\gamma\,n_t}{\kappa^2}\,(dy_t - \gamma\,\hat{e}_t\,dt), \qquad \hat{e}_0 = \mathbb{E}(e_0),$$

and m_t, n_t are the solutions of the equations

$$\frac{dm_t}{dt} - \frac{\beta^2}{Q}\,m_t^2 + P = 0, \qquad \frac{dn_t}{dt} = \sigma^2 - \frac{\gamma^2}{\kappa^2}\,n_t^2,$$

with $m_T = 0$ and $n_0 = \mathrm{var}(e_0)$. A numerical simulation is shown in figure 7.3.

Remark 7.5.5. Though we have only considered the finite time horizon cost, it is not difficult to develop also the time-average and the discounted versions of the LQG problem; see, e.g., [Dav77]. In fact, the usefulness of the separation principle in the linear setting is not restricted to quadratic costs; we may choose the running and terminal costs essentially arbitrarily, and the optimal control will still be expressible as a function of the Kalman-Bucy filter [FR75] (though, like in the completely observed case, there is no analytic solution for the feedback function when the cost is not quadratic). The quadratic cost has a very special property, however, that is not shared by other cost functions. Note that for the quadratic cost, the optimal feedback function for the partially observed case $u_t^* = \alpha(t, \hat{X}_t)$ is the same function as in the completely observed case $u_t^* = \alpha(t, X_t)$, where we have merely replaced the system state by its estimate! This is called *certainty equivalence*. Though the optimal control problem still separates for linear systems with non-quadratic cost, certainty equivalence no longer holds in that case. In other words, in the latter case we still have $u_t^* = \alpha(t, \hat{X}_t)$, but for the completely observed problem $u_t^* = \alpha'(t, X_t)$ with $\alpha' \neq \alpha$.

7.6 Transmitting a message over a noisy channel

We conclude this chapter with a nice control example from communication theory which does not quite fall within our standard control framework: the transmission of a message over a noisy channel with noiseless feedback. The problem was briefly described in the Introduction, but let us recall the basic setting here. We will be content to treat only the simplest setting and to prove optimality within a restricted class of strategies; more general results can be found in [LS01b, section 16.4].

Two parties—a *transmitter* and a *receiver*—are connected through a noisy communication channel. This means that when the transmitter sends the signal u_t through the channel, the receiver observes the noisy signal $y_t = u_t + \xi_t$ where ξ_t is white noise. The transmitter cannot just send any signal u_t, however. First, we have a *time constraint:* the transmitter only has access to the channel in a fixed time interval $[0, T]$. Second, the transmitter has a *power constraint:* it can only send signals which satisfy

$$\mathbb{E}\left[\frac{1}{t}\int_0^t (u_s)^2\, ds\right] \leq P \qquad \forall\, t \in [0, T],$$

where P bounds the signal power per unit time. On the other hand, we will presume that the receiver may send a response to the transmitter in a noiseless manner, i.e., that there is a noiseless feedback channel. This setup is illustrated in figure 7.4.

Let us now turn to the message. We will investigate the simplest type of message: the transmitter has obtained a single Gaussian random variable θ, which is $\mathcal{F}_0$-measurable and thus independent of B_t, to transmit to the receiver. We are thus faced with the following problem: we would like to optimize our usage of the communication channel by choosing wisely the encoding strategy employed by the transmitter, the decoding strategy employed by the receiver, and the way in which the receiver and transmitter make use of the feedback channel, so that the receiver can form the best possible estimate of θ at the end of the day given the time and power constraints.

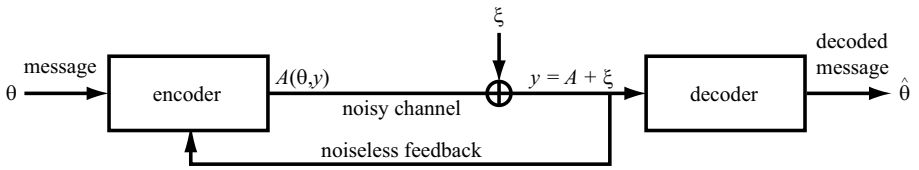

Figure 7.4. Figure 0.4 revisited: setup for the transmission of a message over a noisy channel with noiseless feedback. Further details can be found in the text.

As usual, we will work with the integrated observations,

$$Y_t^u = \int_0^t u_s\,ds + B_t, \qquad \mathcal{F}_t^{R,u} = \sigma\{Y_s^u : s \le t\}, \qquad \mathcal{F}_t^{T,u} = \sigma\{\theta, \mathcal{F}_t^{R,u}\},$$

where B_t is an $\mathcal{F}_t$-Wiener process on the probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, \mathbb{P})$. As the transmitter has access to Y_t^u (by virtue of the noiseless feedback, the receiver can always forward the received signal Y_t^u back to the transmitter), we may choose u_t to be $\mathcal{F}_t^{T,u}$-adapted. On the other hand, the receiver will only have access to the noisy observations $\mathcal{F}_t^{R,u}$, and at the end of the communication period the receiver will form an $\mathcal{F}_T^{R,u}$-measurable estimate $\hat{\theta}^u$ of the message θ. We now have two problems:

- The transmitter must choose the optimal encoding strategy u.
- The receiver must choose the optimal decoding strategy $\hat{\theta}^u$.

By "optimal", we mean that we wish to minimize the mean square error $J'[u, \hat{\theta}^u] = \mathbb{E}((\theta - \hat{\theta}^u)^2)$ over all encoding/feedback strategies u and all decoding strategies $\hat{\theta}^u$.

The second problem—the choice of an optimal decoding strategy—is easy to solve. After all, we know that at time t, the best estimate of θ based on the observation history is given by $\hat{\theta}_t^u = \mathbb{E}(\theta|\mathcal{F}_t^{R,u})$, regardless of the encoding strategy u employed by the receiver. In principle, this solves one half of the problem.

In practice, if we choose a complicated encoding strategy, the resulting decoder (filter) may be very complicated; in particular, it will be difficult to find the optimal encoding strategy when we have so much freedom. We will therefore restrict ourselves to a particularly simple class of encoders, which can be written as $u_t = a_t + b_t\theta$ where a_t is $\mathcal{F}_t^{R,u}$-adapted and b_t is non-random. We will seek an optimal encoding strategy within this class of *linear* encoding strategies. The advantage of the linear strategies is that the resulting filters $\hat{\theta}_t^u$ are easily computable; let us begin by doing this.

Lemma 7.6.1. *Consider the linear encoding strategy $u_t = a_t + b_t\theta$, and define the simplified strategy $u'_t = b_t\theta$. If $\mathcal{F}_t^{R,u} = \mathcal{F}_t^{R,u'}$ for all $t \in [0,T]$, then*

$$d\hat{\theta}_t^u = \hat{P}_t^u b_t\,(dY_t^u - a_t\,dt - b_t\hat{\theta}_t^u\,dt), \qquad \frac{d\hat{P}_t^u}{dt} = -(b_t\hat{P}_t^u)^2,$$

where $\hat{\theta}_0^u = \mathbb{E}(\theta)$ and $\hat{P}_0^u = \mathrm{var}(\theta)$.

Proof. For u', we simply obtain a Kalman-Bucy filter with $X_0 = \theta$, $A(t) = B(t) = C(t) = 0$, $H(t) = b_t$, and $K(t) = 1$. But $\hat{\theta}^u_t = \hat{\theta}^{u'}_t$ follows trivially from the assumption on the equality of the σ-algebras, so we obtain the result. □

We can now find the optimal strategy within this class.

Proposition 7.6.2. *Within the class of linear encoding strategies which satisfy the condition of lemma 7.6.1 and the power constraint, an optimal strategy u^* is given by*

$$u^*_t = \sqrt{\frac{P}{\operatorname{var}(\theta)}}\, e^{Pt/2} \left(\theta - \hat{\theta}^{u^*}_t\right).$$

The ultimate mean square error for this strategy is $\mathbb{E}((\theta - \hat{\theta}^{u^*}_T)^2) = \operatorname{var}(\theta)\, e^{-PT}$.

Proof. Let u be any linear strategy which satisfies the required conditions. The trick is to find a lower bound on the estimation error given the power constraint on the signal. The problem then reduces to seeking a strategy that attains this lower bound and satisfies the power constraint.

Note that for the strategy u, the mean square estimation error is precisely $\hat{P}^u_T$. Now compute

$$\frac{d}{dt}\log(\hat{P}^u_t) = -b_t^2 \hat{P}^u_t \quad \Longrightarrow \quad \hat{P}^u_t = \operatorname{var}(\theta)\, \exp\left(-\int_0^t b_s^2 \hat{P}^u_s\, ds\right).$$

On the other hand, note that

$$\mathbb{E}((u_t)^2) = \mathbb{E}((a_t + b_t\hat{\theta}^u_t + b_t(\theta - \hat{\theta}^u_t))^2) = \mathbb{E}((a_t + b_t\hat{\theta}^u_t)^2) + b_t^2 \hat{P}^u_t \geq b_t^2 \hat{P}^u_t,$$

where we have used the properties of the conditional expectation to conclude that we may set $\mathbb{E}((a_t + b_t\hat{\theta}^u_t)(\theta - \hat{\theta}^u_t)) = 0$. But then our power constraint requires that

$$\int_0^t b_s^2 \hat{P}^u_s\, ds \leq \mathbb{E}\left[\int_0^t (u_s)^2\, ds\right] \leq Pt,$$

so we conclude that $\hat{P}^u_t \geq \operatorname{var}(\theta)\, e^{-Pt}$. To find a strategy that attains this bound, note that

$$\frac{d}{dt}\operatorname{var}(\theta)\, e^{-Pt} = -P\operatorname{var}(\theta)\, e^{-Pt} = -\left(\operatorname{var}(\theta)\, e^{-Pt}\sqrt{\frac{P\, e^{Pt}}{\operatorname{var}(\theta)}}\right)^2,$$

so $b_t = e^{Pt/2}\sqrt{P/\operatorname{var}(\theta)}$ gives $\hat{P}^u_t = \operatorname{var}(\theta)\, e^{-Pt}$. Thus we must choose this b_t to obtain any optimal strategy, provided we can find an a_t such that the resulting strategy satisfies the power constraint. But for this choice of b_t, we find that

$$\mathbb{E}\left[\int_0^t (u_s)^2\, ds\right] = \mathbb{E}\left[\int_0^t (a_s + b_s\hat{\theta}^u_s)^2\, ds\right] + \int_0^t b_s^2 \hat{P}^u_s\, ds = \mathbb{E}\left[\int_0^t (a_s + b_s\hat{\theta}^u_s)^2\, ds\right] + Pt,$$

so the power constraint is satisfied if and only if $a_t + b_t\hat{\theta}^u_t = 0$ for all t. This yields the strategy u^*_t. It remains to check that the strategy u^*_t satisfies the condition of lemma 7.6.1; but this is easily done following the same logic as in the proof of proposition 7.3.9. □

Have we gained anything by using the feedback channel? Let us see what happens if we disable the feedback channel; in this case, a_t can no longer depend on the observations and is thus also non-random. We now obtain the following result.

Proposition 7.6.3. *Within the class of linear encoding strategies without feedback which satisfy the power constraint, an optimal strategy u^* is given by*

$$u_t^* = \sqrt{\frac{P}{\operatorname{var}(\theta)}}\,(\theta - \mathbb{E}(\theta)).$$

The ultimate mean square error for this strategy is $\mathbb{E}((\theta - \hat{\theta}_T^{u^*})^2) = \operatorname{var}(\theta)/(1+PT)$.

Proof. This is the same idea as in the previous proof, only we now require that a_t is non-random (note that in this case the condition of lemma 7.6.1 is automatically satisfied). The equation for $\hat{P}_t^u$ can be solved explicitly: it is easily verified that

$$\hat{P}_t^u = \frac{\operatorname{var}(\theta)}{1 + \operatorname{var}(\theta)\int_0^t (b_s)^2\,ds}.$$

On the other hand, note that

$$\mathbb{E}((u_t)^2) = \mathbb{E}((a_t + b_t\,\mathbb{E}(\theta) + b_t(\theta - \mathbb{E}(\theta)))^2) = \mathbb{E}((a_t + b_t\,\mathbb{E}(\theta))^2) + (b_t)^2\operatorname{var}(\theta).$$

Then we obtain, using the power constraint,

$$\operatorname{var}(\theta)\int_0^t (b_s)^2\,ds \le \mathbb{E}\left[\int_0^t (u_s)^2\,ds\right] \le Pt \quad\Longrightarrow\quad \hat{P}_t^u \ge \frac{\operatorname{var}(\theta)}{1+Pt}.$$

The remainder of the proof follows easily. □

Remark 7.6.4. Evidently the strategy that uses the feedback channel performs much better than the strategy without feedback. It is illuminating in this regard to investigate the particular form of the optimal strategies. Note that in the absence of the power constraint, we would have no problem sending the message across the noisy channel; we could just transmit θ directly over the channel with some large gain factor, and by cranking up the gain we can make the signal to noise ratio arbitrarily large (and thus the estimation error arbitrarily small). However, with the power constraint in place, we have to choose wisely which information we wish to spend our power allowance on. Clearly it is not advantageous to waste power in transmitting something that the receiver already knows; hence the optimal strategies, rather than transmitting the message itself, try to transmit the discrepancy between the message and the part of the message thas is known to the receiver. Here feedback is of great help: as the transmitter knows what portion of the message was received on the other end, it can spend its remaining power purely on transmitting the parts of the message that were corrupted (it does this by only transmitting the discrepancy between the message and the receiver's estimate of the message). On the other hand, the feedbackless transmitter has no idea what the receiver knows, so the best it can do is subtract from θ its mean (which is assumed to be known both to the transmitter and to the receiver).

Surprisingly, perhaps, these results are not restricted to the linear case; in fact, it turns out that the encoding strategy of proposition 7.6.2 is optimal even in the class of all nonlinear encoding strategies. It would be difficult to prove this directly, however, as this would require quantifying the mean-square error for a complicated set of

nonlinear filters. Instead, such a claim is proved by *information-theoretic* arguments, which are beyond our scope. The idea of the proof is still the same: we seek to obtain a lower bound on the mean-square error given the power constraint, and show that our candidate strategy (the linear strategy of proposition 7.6.2) attains this bound. However, techniques from information theory can be used to obtain generic lower bounds on the mean-square error of an estimator which are not specific to a particular type of filter, so that the complications of the nonlinear case can be avoided. Nonetheless the filtering theory is crucial in order to demonstrate that the optimal strategy attains the lower bound, and to give an explicit expression for the estimator (which we have already done). Further details can be found in [LS01b, section 16.4], as well as an extension of these results to more complicated (time-dependent) messages.

7.7 Further reading

There are two main approaches to the nonlinear filtering problem. The first is the *reference probability method* which we have used in this chapter. The second approach, the *innovations method*, runs almost in the opposite direction. There one begins by proving that the innovations process is a Wiener process. Then, using a martingale representation argument (with some delicate technicalities), one can prove that the filter can be expressed as the sum of a time integral and a stochastic integral with respect to the innovations. It then remains, using some clever tricks, to identify the integrands.

Both approaches have turned out to be extremely fruitful in various situations. As you likely realize by this point, an unpleasant feature of the reference probability method is that in many cases the Girsanov change of measure Λ_t is not square-integrable, so that we can not apply a result such as lemma 7.2.7. The result of a systematic application of the reference probability method can be seen in the book by Bensoussan [Ben92]: there is a constant need to perform truncation and limiting arguments to circumvent the technical problems. Not quite as detailed, but more elegant, are the excellent lecture notes by Pardoux [Par91]. This is a great place to start reading about the reference probability method (if you read French). A somewhat different point of view and setting can be found in Elliott, Aggoun and Moore [EAM95].

The innovations method, which we have not developed here, has less trouble with the sort of technical issues that the reference probability method suffers from, and is more convenient when there are correlations between the noise that is driving the signal process and the observation noise (a case which we have not considered). A very accessible introduction to the innovations approach is the book by Krishnan [Kri05], which is highly recommended. The bible of the innovations approach remains the two-volume extravaganza by Liptser and Shiryaev [LS01a, LS01b], while the book by Kallianpur [Kal80] provides another in-depth treatment. A nice discussion of the innovations approach also appears in the book by Elliott [Ell82].

Both approaches, and much more besides (including discrete time filtering), are treated in a wonderful set of lecture notes by Chigansky [Chi04].

The problem of making sense of the nonlinear filtering equations, such as the Zakai equation, as stochastic PDEs, is treated in various places. Good places to look are Kunita [Kun90], Bensoussan [Ben92], Pardoux [Par82, Par91], and the book by

Rozovskii [Roz90]. The issue of efficient numerical algorithms is another story; many references were already mentioned in the chapter, but see in particular the recent review [Cri02]. There are various interesting issues concerning the finite-dimensional realization of filtering problems; the volumes [HW81, Mit82] contain some interesting articles on this topic. Another very useful topic which we have overlooked, the *robust* or *pathwise* definition of the filter, is discussed in an article by Davis in [HW81].

The Kalman-Bucy filter is treated in detail in many places; see, e.g., the book by Davis [Dav77] and Liptser and Shiryaev [LS01a, LS01b]. Our treatment, through stochastic control, was heavily inspired by the treatment in Fleming and Rishel [FR75] and in Bensoussan [Ben92]. The relations between filtering and stochastic control go very deep indeed, and are certainly not restricted to the linear setting; on this topic, consult the beautiful article by Mitter and Newton [MN03]. The Kalman-Bucy filter can be extended also to the general conditionally Gaussian case where $A(t)$, $B(t)$, $C(t)$, $H(t)$ and $K(t)$ are all adapted to the observations, see Liptser and Shiryaev [LS01b], as well as to the case where X_0 has an arbitrary distribution (i.e., it is non-Gaussian); for the latter, see the elegant approach by Makowski [Mak86].

The Shiryaev-Wonham filter is due to Shiryaev [Shi63, Shi73] and, in a more general setting which allows the signal to be an arbitrary finite state Markov process, due to Wonham [Won65]. Our treatment was inspired by Rogers and Williams [RW00b].

On the topic of partially observed control, Bensoussan [Ben92] is a good source of information and further references. Our treatment was inspired by Fleming and Rishel [FR75], which follows closely the original article by Wonham [Won68b] (for results in the finite state setting see Segall [Seg77] and Helmes and Rishel [HR92]). Finally, the transmission of a message through a noisy channel, and many other applications, are treated in the second volume of Liptser and Shiryaev [LS01b].

CHAPTER 8

Optimal Stopping and Impulse Control

In the previous chapters, we have discussed several control problems where the goal was to optimize a certain performance criterion by selecting an apppropriate feedback control policy. In this chapter, we will treat a somewhat different set of control problems; rather than selecting a continuous control to be applied to an auxiliary input in the system equations, our goal will be to select an optimal *stopping time* to achieve a certain purpose. Such problems show up naturally in many situations where a timing decision needs to be made, e.g., when is the best time to sell a stock? When should we decide to bring an apparatus, which may or may not be faulty, in for repair (and pay the repair fee)? How long do we need to observe an unknown system to be able to select one of several hypotheses with sufficient confidence? Such problems are called *optimal stopping* problems, and we will develop machinery to find the optimal stopping times. These ideas can also be extended to find optimal control strategies in which feedback is applied to the system at a discrete sequence of times; we will briefly discuss such *impulse control* problems at the end of this chapter.

8.1 Optimal stopping and variational inequalities

The optimal stopping problem

As usual, we work on $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, \mathbb{P})$ with an m-dimensional $\mathcal{F}_t$-Wiener process W_t, and we will describe the system of interest by the stochastic differential equation

$$dX_t = b(X_t)\,dt + \sigma(X_t)\,dW_t \qquad (X_t \in \mathbb{R}^n),$$

where X_0 is $\mathcal{F}_0$-measurable and $b : \mathbb{R}^n \to \mathbb{R}^n$, $\sigma : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ satisfy appropriate conditions that ensure existence and uniqueness of the solution.

Remark 8.1.1. As with the infinite time costs in chapter 6, we will find it convenient to choose b and σ to be time-independent. However, time can always be added simply by considering the $(n+1)$-dimensional system $X'_t = (X_t, t)$.

For an $\mathcal{F}_t$-stopping time τ, define the cost functional

$$J[\tau] = \mathbb{E}\left[\int_0^\tau e^{-\lambda s}\, w(X_s)\, ds + e^{-\lambda\tau}\, z(X_\tau)\right],$$

where $\lambda \geq 0$ is a discount factor (the case $\lambda = 0$ corresponds to the non-discounted setting). We will call τ *admissible* if $J[\tau]$ is well defined; any stopping time will be admissible, e.g., if w, z are nonnegative (or $\lambda > 0$ and w, z are bounded). The goal of the *optimal stopping problem* is to find a stopping time τ^* which minimizes $J[\tau]$. In principle, the optimal stopping time τ^* can be an arbitrarily complicated functional of the sample paths of X_t. However, there is a special type of stopping time which plays the same role in optimal stopping theory as did Markov strategies in chapter 6: this is precisely the class of stopping rules τ which are the first exit time of X_t from some set $D \subset \mathbb{R}^n$, i.e., $\tau = \inf\{t : X_t \notin D\}$. The set D is then called the *continuation region* for the stopping rule τ. Conveniently, it turns out that optimal stopping rules are of this form, just like optimal control strategies turn out to be Markov.

A heuristic calculation

As in chapter 6, we will mostly concentrate on obtaining a useful verification theorem. However, to clarify where the equations in the verification theorem come from, it is helpful to first obtain the appropriate equations in a heuristic manner. Let us do this now. *We will disregard any form of technical precision until further notice.*

Let τ be an admissible stopping rule. We define the *cost-to-go* $J^\tau(x)$ as

$$J^\tau(X_0) = \mathbb{E}\left[\int_0^\tau e^{-\lambda s}\, w(X_s)\, ds + e^{-\lambda\tau}\, z(X_\tau)\,\middle|\, X_0\right].$$

Note that $J^\tau(x)$ is the cost of the stopping rule τ when $X_0 = x$ is non-random.

Now suppose, as we did in the corresponding discussion in chapter 6, that there is a stopping rule τ^*, with continuation region D, which minimizes $J^\tau(x)$ for every x, and define the *value function* as the optimal cost-to-go $V(x) = J^{\tau^*}(x)$. We will try to find an equation for $V(x)$. To this end, let τ be any admissible stopping rule, and define the new stopping rule $\tau' = \inf\{t \geq \tau : X_t \notin D\}$. Then τ' is the rule under which we do not stop until the time τ, and continue optimally afterwards, and by assumption $J[\tau^*] \leq J[\tau']$. But then it is not difficult to see that

$$V(X_0) \leq J^{\tau'}(X_0) = \mathbb{E}\left[\int_0^\tau e^{-\lambda s}\, w(X_s)\, ds + e^{-\lambda\tau}\, V(X_\tau)\,\middle|\, X_0\right],$$

where we have used the stong Markov property of X_t (see remark 5.2.3) and the tower property of the conditional expectation. On the other hand, we obtain an equality in this expression, rather than an inequality, if we choose $\tau \leq \tau^*$ (why?).

Now suppose that $V(x)$ is sufficiently smooth to apply Itô's rule. Then

$$e^{-\lambda\tau}\,V(X_\tau) = V(X_0) + \int_0^\tau e^{-\lambda s}\{\mathscr{L}V(X_s) - \lambda\,V(X_s)\}\,ds + \int_0^\tau \cdots\,dW_s.$$

If additionally the expectation of the last term vanishes, then

$$V(X_0) = \mathbb{E}\left[\int_0^\tau e^{-\lambda s}\{\lambda\,V(X_s) - \mathscr{L}V(X_s)\}\,ds + e^{-\lambda\tau}\,V(X_\tau)\,\middle|\,X_0\right].$$

First, suppose that $\tau \le \tau^*$. Then we obtain the equality

$$0 = \mathbb{E}\left[\int_0^\tau e^{-\lambda s}\{\mathscr{L}V(X_s) - \lambda V(X_s) + w(X_s)\}\,ds\,\middle|\,X_0\right].$$

As this holds for any $\tau \le \tau^*$, we must have $\mathscr{L}V(X_0) - \lambda V(X_0) + w(X_0) = 0$ *provided* $\tau^* > 0$, i.e., for $X_0 \in D$; for $X_0 \notin D$ this expression is identically zero, and we do not learn anything from it! But for $x \notin D$, clearly $V(x) = z(x)$; hence

$$\mathscr{L}V(x) - \lambda V(x) + w(x) = 0 \quad \text{for } x \in D, \qquad V(x) = z(x) \quad \text{for } x \notin D.$$

Now consider the case that τ is arbitrary. Proceeding in the same way as above, we obtain $\mathscr{L}V(x) - \lambda V(x) + w(x) \ge 0$; on the other hand $J[\tau^*] \le J[0]$ (we can do at least as well as stopping immediately), so that in particular $V(x) \le z(x)$. Hence

$$\mathscr{L}V(x) - \lambda V(x) + w(x) \ge 0, \qquad V(x) \le z(x) \qquad \text{for all } x.$$

Evidently $V(x)$ satisfies the following equation:

$$\min\{\mathscr{L}V(x) - \lambda V(x) + w(x),\ z(x) - V(x)\} = 0.$$

This is not a PDE in the usual sense; it is called a *variational inequality*. Just like in the optimal control case, where the Bellman equation reduces the optimization problem to a pointwise minimization over all possible control actions, the variational inequality reduces the optimal stopping problem to a pointwise minimization over our two options: to continue, or to stop. It is important to note that if $V(x)$ is a unique solution to the variational inequality, then we can completely reconstruct the continuation region D: it is simply $D = \{x : V(x) < z(x)\}$. Hence it suffices, as in the optimal control setting, to solve a (nonlinear, variational inequality) PDE for the value function, in order to be able to construct the optimal strategy.

Remark 8.1.2. There are much more elegant treatments of the optimal stopping theory which can be made completely rigorous with some effort. One of these methods, due to Snell, is closely related to the martingale dynamic programming principle in optimal control (see remark 6.1.7) and works in a general setting. Another method, due to Dynkin, characterizes the optimal cost of stopping problems for the case where X_t is a Markov process. Both these methods are extremely fundamental to optimal stopping theory and are well worth studying; see, e.g., [PS06].

A verification theorem

The previous discussion is only intended as motivation. We have made various entirely unfounded assumptions, *which you should immediately discard from this point onward*. Rather than making the story above rigorous, we proceed, as in chapter 6, in the opposite direction: we will assume that we have found a sufficiently smooth solution to the appropriate variational inequality, and prove a verification theorem that guarantees that this solution does indeed give rise to an optimal stopping rule.

Proposition 8.1.3. *Let $K \subset \mathbb{R}^n$ be a set such that $X_t \in K$ for all t. Suppose there is a function $V : K \to \mathbb{R}$, which is sufficiently smooth to apply Itô's rule, such that*

$$\min\{\mathscr{L}V(x) - \lambda V(x) + w(x),\, z(x) - V(x)\} = 0,$$

and $|\mathbb{E}(V(X_0))| < \infty$. Define the set $D = \{x \in K : V(x) < z(x)\}$, and denote by $\mathfrak{K}$ the class of admissible stopping rules τ such that $\tau < \infty$ a.s. and

$$\mathbb{E}\left[\sum_{i=1}^{n}\sum_{k=1}^{m}\int_0^\tau e^{-\lambda s}\,\frac{\partial V}{\partial x^i}(X_s)\,\sigma^{ik}(X_s)\,dW_s^k\right] = 0.$$

Suppose that $\tau^* = \inf\{t : X_t \notin D\}$ is in $\mathfrak{K}$. Then $J[\tau^*] \le J[\tau]$ for any $\tau \in \mathfrak{K}$, and the optimal cost can be expressed as $\mathbb{E}(V(X_0)) = J[\tau^*]$.

Remark 8.1.4. Often $K = \mathbb{R}^n$, but we will use the more general statement later on.

Proof. Applying Itô's rule to $e^{-\lambda t}V(X_t)$ and using the condition on strategies in $\mathfrak{K}$, we obtain

$$\mathbb{E}(V(X_0)) = \mathbb{E}\left[\int_0^\tau e^{-\lambda s}\{\lambda V(X_s) - \mathscr{L}V(X_s)\}\,ds + e^{-\lambda\tau}\,V(X_\tau)\right]$$

for $\tau \in \mathfrak{K}$. But the variational inequality implies $V(x) \le z(x)$ and $\lambda V(x) - \mathscr{L}V(x) \le w(x)$, so we find that $\mathbb{E}(V(X_0)) \le J[\tau]$. On the other hand, for $\tau = \tau^*$, we these inequalities become equalities, so we find that $J[\tau^*] = \mathbb{E}(V(X_0)) \le J[\tau]$. This establishes the claim. □

In this verification result, we required that $V(x)$ "is sufficiently smooth to apply Itô's rule". It would seem that we should just assume that V is in C^2, as this is the requirement for Itô's rule. Unfortunately, hardly any optimal stopping problem gives rise to a value function in C^2. The problems occur on the boundary ∂D of D: often $V(x)$ is C^2 on $K\backslash\partial D$, but on ∂D it is only C^1. We thus need to extend Itô's rule to this situation. There are various technical conditions under which this is possible; the most elegant is the following result, which holds only in one dimension.

Proposition 8.1.5 (Relaxed Itô rule in one dimension). *Suppose that $V : \mathbb{R} \to \mathbb{R}$ is C^1 and admits a (not necessarily continuous) second derivative in the sense that there exists a measurable function $\partial^2 V/\partial x^2$ such that*

$$\frac{\partial V}{\partial x}(x) - \frac{\partial V}{\partial x}(0) = \int_0^x \frac{\partial^2 V}{\partial x^2}(y)\,dy, \qquad x \in \mathbb{R}.$$

Then Itô's rule still applies to $V(X_t)$.

For the proof of this statement, see [RW00b, lemma IV.45.9]. The proof is a little too difficult to us; it requires, in essence, to show that X_t does not spend any time at the discontinuity points of $\partial^2 V/\partial x^2$ (i.e., the amount of time spent at the discontinuity points has Lebesgue measure zero). For generalizations to the multidimensional case, see [Kry80, section 2.10] and particularly [PS93] (see also [Øks03, theorem 10.4.1] and [Fri75, theorem 16.4.1]). For our purposes proposition 8.1.5 will suffice, as we will restrict ourselves to one-dimensional examples for simplicity.

Remark 8.1.6 (The principle of smooth fit). The fact that $V(x)$ is generally not C^2 is not surprising; on the other hand, the fact that $V(x)$ *should* be C^1 is not at all obvious! Nonetheless, in many cases it can be shown that the gradient of $V(x)$ does indeed need to be continuous on the boundary ∂D; this is called the *principle of smooth fit* (see, e.g., [PS06, DK03] for proofs). This turns out to be an extremely useful tool for finding an appropriate solution of the variational inequality. In general, there are many solutions to the variational inequality, each leading to a different continuation set D; however, it is often the case that only one of these solutions is continuously differentiable. Only this solution, then, satisfies the conditions of the verification theorem, and thus the principle of smooth fit has helped us find the correct solution to the optimal stopping problem. We will shortly see this procedure in action.

Let us treat an interesting example (from [Øks03]).

Example 8.1.7 (Optimal resource extraction). We are operating a plant that extracts natural gas from an underground well. The total amount of natural gas remaining in the well at time t is denoted R_t (so the total amount of extracted natural gas is $R_0 - R_t$). Moreover, the rate at which we can extract natural gas from the well is proportional to the remaining amount: that is, when the plant is in operation, the amount of natural gas in the well drops according the equation

$$\frac{d}{dt} R_t = -\lambda R_t,$$

where $\lambda > 0$ is the proportionality constant. After the gas has been extracted, it is sold on the market at the current market price P_t, which is given by the equation

$$dP_t = \mu P_t \, dt + \sigma P_t \, dW_t,$$

where $\mu > 0$. However, it costs money to operate the plant: in order to keep the plant running we have to pay K dollars per unit time. The total amount of money made by time t by extracting natural gas and selling it on the market is thus given by

$$\int_0^t P_s \, d(R_0 - R_s) - Kt = \int_0^t (\lambda R_s P_s - K) \, ds.$$

It seems inevitable that at some point in time it will no longer be profitable to keep the plant in operation: we will not be able to extract the natural gas sufficiently rapidly to be able to pay for the operating costs of the plant. We would like to determine when

would be the best time to call it quits, i.e., we would like to find a stopping time τ^* which maximizes the expected profit $-J[\tau]$ up to time τ. We thus seek to minimize

$$J[\tau] = \mathbb{E}\left[\int_0^\tau (K - \lambda R_s P_s)\, ds\right].$$

This is precisely an optimal stopping problem of the type we have been considering.

The problem can be simplified by noting that the cost functional depends only on the quantity $S_t = R_t P_t$. Using Itô's rule, we find that

$$dS_t = (\mu - \lambda) S_t\, dt + \sigma S_t\, dW_t, \qquad J[\tau] = \mathbb{E}\left[\int_0^\tau (K - \lambda S_s)\, ds\right].$$

As $S_t \geq 0$ a.s. for all t, we can apply proposition 8.1.3 with $K = [0, \infty[$. The variational inequality for this problem can be written as

$$\min\left\{\frac{\sigma^2 x^2}{2}\frac{\partial^2 V(x)}{\partial x^2} + (\mu - \lambda)x\,\frac{\partial V(x)}{\partial x} + K - \lambda x,\ -V(x)\right\} = 0.$$

Thus on D^c, we must have $V(x) = 0$ and $\mathscr{L}V(x) + K - \lambda x = K - \lambda x \geq 0$; in particular, if $x \in D^c$, then $x \leq K/\lambda$, so we conclude that $]K/\lambda, \infty[\subset D$. Let us now try to solve for $V(x)$ on D. To this end, consider the PDE

$$\frac{\sigma^2 x^2}{2}\frac{\partial^2 V(x)}{\partial x^2} + (\mu - \lambda)x\,\frac{\partial V(x)}{\partial x} + K - \lambda x = 0.$$

Let us try a solution of the form

$$V_c(x) = -\frac{K\log(x)}{\mu - \lambda - \sigma^2/2} + \frac{\lambda\, x}{\mu - \lambda} + c.$$

If $V(x) = V_c(x)$ on D, then it must be that $V_c(x) < 0$ on $]K/\lambda, \infty[$; in particular, this means that we must require that $\mu < \lambda$. Intuitively this makes sense: if the price of natural gas were to grow at a faster rate than the rate at which we deplete our well, then it would always pay off to keep extracting more natural gas!

Let us thus assume that $\mu < \lambda$, and we are seeking a solution of the form $V(x) = V_c(x)$ on D and $V(x) = 0$ on D^c. To determine the appropriate c and D, we will try to paste the solutions $V_c(x)$ and 0 together in such a way that the result is C^1—i.e., we are going to use the principle of smooth fit. To this end, note that the derivative of V must vanish on the boundary of D (as $V(x) = 0$ on D^c). But

$$\frac{dV_c(x)}{dx} = -\frac{Kx^{-1}}{\mu - \lambda - \sigma^2/2} + \frac{\lambda}{\mu - \lambda} = 0 \quad \Longrightarrow \quad x = \frac{K}{\lambda}\frac{\mu - \lambda}{\mu - \lambda - \sigma^2/2} \equiv x^*.$$

Thus D must be of the form $]x^*, \infty[$ (note that $x^* < K/\lambda$, so this indeed makes sense). On the other hand, $V(x)$ must be continuous at x^*, so if $V(x) = V_c(x)$ we should have $V_c(x^*) = 0$. This allows us to select the appropriate value c^* of c:

$$V_{c^*}(x^*) = 0 \quad \Longrightarrow \quad c^* = \frac{K\log(x^*)}{\mu - \lambda - \sigma^2/2} - \frac{\lambda\, x^*}{\mu - \lambda}.$$

We have thus shown that the variational inequality is solved by the value function $V(x) = 0$ for $x \leq x^*$, and $V(x) = V_{c^*}(x)$ for $x > x^*$; note that $V(x)$ is C^1 on $[0, \infty[$ and C^2 on $[0, \infty[\backslash\{x^*\}$. Our candidate stopping rule is thus $\tau^* = \inf\{t : X_t \leq x^*\}$.

To conclude that τ^* is indeed optimal, it remains to show that $\tau^* \in \mathfrak{K}$. This is indeed possible whenever $\mu < \lambda$ using a more refined version of the optional stopping theorem than we have discussed; see [RW00a, theorems II.69.2 and II.77.5]. For sake of simplicity, let us verify that $\tau^* \in \mathfrak{K}$ under the more restrictive assumption that $\mu - \lambda + \sigma^2/2 < 0$. Recall that we must assume that $\mathbb{E}(V(S_0))$ is finite, and we will also assume without loss of generality that $S_0 \geq x^*$ a.s.

First we will establish that $\mathbb{E}(\tau^*) < \infty$. To this end, note that

$$\log(S_{t\wedge\tau^*}) = \log(S_0) + (\mu - \lambda - \sigma^2/2)\, t \wedge \tau^* + \sigma\, W_{t\wedge\tau^*}.$$

As $\mathbb{E}(V(S_0))$ is finite, $\mathbb{E}(\log(S_0))$ is finite also, and we find that $\mathbb{E}(\log(S_{t\wedge\tau^*})) = \mathbb{E}(\log(S_0)) + (\mu - \lambda - \sigma^2/2)\,\mathbb{E}(t \wedge \tau^*)$. In particular, by monotone convergence,

$$\mathbb{E}(\tau^*) = \frac{\mathbb{E}(\log(S_0)) - \lim_{t\to\infty} \mathbb{E}(\log(S_{t\wedge\tau^*}))}{\sigma^2/2 + \lambda - \mu}.$$

But $\mathbb{E}(\log(S_{t\wedge\tau^*})) \geq \log x^*$, so $\mathbb{E}(\tau^*) < \infty$. Next, we need to show that

$$\mathbb{E}\left[\int_0^{\tau^*} \frac{\partial V}{\partial x}(S_s)\,\sigma\, S_s\, dW_s\right] = \mathbb{E}\left[\int_0^{\tau^*} (C_1\, S_s + C_2)\, dW_s\right] = 0,$$

where C_1 and C_2 are the appropriate constants. The integral over C_2 has zero expectation by lemma 6.3.4. To deal with the integral over S_t, note that for $m < n$

$$\mathbb{E}\left[\left(\int_{m\wedge\tau^*}^{n\wedge\tau^*} S_s\, dW_s\right)^2\right] = \mathbb{E}\left[\int_{m\wedge\tau^*}^{n\wedge\tau^*} (S_s)^2\, ds\right] \leq \mathbb{E}\left[\int_0^{\infty} (S_s)^2\, ds\right].$$

But you can verify using Itô's rule that the term on the right is finite whenever we have $\mu - \lambda + \sigma^2/2 < 0$. Hence if this is the case, we find using dominated convergence

$$\int_0^{n\wedge\tau^*} S_s\, dW_s \to \int_0^{\tau^*} S_s\, dW_s \quad \text{in } \mathcal{L}^2 \quad \Longrightarrow \quad \mathbb{E}\left[\int_0^{\tau^*} S_s\, dW_s\right] = 0$$

(use that the integral is a Cauchy sequence in $\mathcal{L}^2$). This is what we set out to show.

It is often useful to be able to introduce an additional constraint in the optimal stopping problem; we would like to find the optimal stopping time *prior* to the time when the system exits a predetermined set K. We will see an example of this below. The corresponding extension of proposition 8.1.3 is immediate, and we omit the proof.

Proposition 8.1.8. *Let $K \subset \mathbb{R}^n$ be a fixed open set, with closure $\overline{K}$ and boundary $\partial K = \overline{K}\backslash K$, and assume that $X_0 \in K$ a.s. Suppose there is a function $V : \overline{K} \to \mathbb{R}$ which is sufficiently smooth to apply Itô's rule, such that $V(x) = z(x)$ on ∂K,*

$$\min\{\mathscr{L}V(x) - \lambda V(x) + w(x),\ z(x) - V(x)\} = 0,$$

and $|\mathbb{E}(V(X_0))| < \infty$. *Define* $D = \{x \in K : V(x) < z(x)\}$, *and let* $\mathfrak{K}$ *be the class of admissible stopping rules* τ *with* $\tau < \infty$ *a.s.,* $\tau \leq \tau_K = \inf\{t : X_t \notin K\}$, *and*

$$\mathbb{E}\left[\sum_{i=1}^{n}\sum_{k=1}^{m}\int_0^\tau e^{-\lambda s}\,\frac{\partial V}{\partial x^i}(X_s)\,\sigma^{ik}(X_s)\,dW_s^k\right] = 0.$$

Suppose that $\tau^* = \inf\{t : X_t \notin D\}$ *is in* $\mathfrak{K}$. *Then* $J[\tau^*] \leq J[\tau]$ *for any* $\tau \in \mathfrak{K}$, *and the optimal cost can be expressed as* $\mathbb{E}(V(X_0)) = J[\tau^*]$.

Note that if K has compact closure, then lemmas 6.3.4 and 6.3.3 can be used to deal with the technical condition, thus avoiding some amount of trench work.

Markov chain approximations

As in the setting of optimal control, most optimal stopping problems do not admit analytic solution. However, as before, Markov chain approximations provide an effective method to approximate the solution of an optimal stopping problem. Let us demonstrate this method through an important example in mathematical finance.

Example 8.1.9 (Optimal exercise for an American put option). We are holding a certain amount of stock, whose price at time t is given by the usual equation

$$dS_t = \mu S_t\,dt + \sigma S_t\,dW_t.$$

At some point in the future we might want to sell our stock, but this is risky: by that point the stock price may have tanked, in which case we would not be able to sell the stock for a reasonable price on the stock market. To mitigate this risk, we may take out a form of insurance on our stock: a *put option*. This is a contract which guarantees that we will be able to sell our stock at some time in the future for a predetermined price K. A *European* put option works precisely in this way: we fix T and K, and the contract guarantees that we may sell our stock for the price K at time T. Hence the payoff from such an option is $(K - S_T)^+$ (because if the stock price is larger than K, we retain the right to sell our stock on the stock market instead).

European put options are not our only choice, however; there are options which allow us more flexibility. In this example, we will investigate an *American* put option. Like in the European case, we fix a price K and a terminal time T. In contrast with the European option, however, an American put option can be exercised at any point in time in the interval $[0, T]$: that is, after purchasing the option at time zero, we may decide at any stopping time $\tau \leq T$ to sell our stock for the price K. If we choose to exercise at time τ, then the payoff from the option is $(K - S_\tau)^+$. The question now becomes: when should we choose to exercise to maximize our payoff from the option? This problem is naturally formulated as an optimal stopping problem.

In general, there is also a bank account involved which gives an interest rate r. It is customary to try to maximize the *discounted* payoff, i.e., we will normalize all our prices by the amount of money we could have made by investing our money in the bank account rather than in the risky stocks. This gives rise to the following optimal

stopping problem: we wish to maximize $\mathbb{E}(e^{-r\tau}(K - S_\tau)^+)$ over all stopping times $\tau \leq T$. In other words, we want to minimize the following cost:

$$J[\tau] = \mathbb{E}(-e^{-r\tau}(K - S_\tau)^+).$$

Remark 8.1.10 (Pricing of American options). An important problem in practice is to determine how an American option should be priced, i.e., how much money should the seller of the option charge for agreeing to issue this contract? Arbitrage pricing theory shows that there is only one fair price for an American option in the current setting; if any other price is charged, either the buyer or the seller of the option can make money for nothing, which is by definition not "fair". Moreover, it turns out that in order to compute the price, we need to solve the above optimal stopping problem for the case $\mu = r$, and the payoff of the optimal stopping rule is then precisely the fair price of the option. The replacement $\mu \mapsto r$ corresponds to a change of measure, which allows us to replicate the payoff of the option by a suitable trading strategy using the martingale representation theorem. The details of this procedure are not difficult, but as they are not directly related to the optimal stopping problem itself we will forgo a discussion here; see, e.g., [Duf01, section 8G]. Suffice it to say that the pricing problem for American options makes the solution of optimal stopping problems of the type which we are considering an important practical problem in real-world finance (though obviously we are here restricting to the very simplest case).

Let us begin by expressing our optimal stopping problem as a variational inequality. As we are seeking stopping times τ which are guaranteed to be less than T, we need to apply proposition 8.1.8 by considering the two-dimensional process $X_t = (t, S_t)$ and the stopping set $K = \{(t, x) : t < T\}$. We thus seek a function $V(t, x)$, defined on the set K, which obeys the following variational inequality:

$$\min\left\{\left(\frac{\partial}{\partial t} + \frac{\sigma^2 x^2}{2}\frac{\partial^2}{\partial x^2} + \mu x \frac{\partial}{\partial x} - r\right) V(t, x),\ -(K - x)^+ - V(t, x)\right\} = 0.$$

If we can find a suitable function $V(t, x)$, then the continuation region for the optimal stopping rule is given by $D = \{(t, x) \in K : V(t, x) + (K - x)^+ < 0\}$. Unfortunately, there is no analytical solution to this problem, so we need to proceed numerically.

Let us define a grid on K. We will split the interval $[0, T]$ into intervals of length $\Delta = T/N$, i.e., we will work with the times $k\Delta$, $k = 0, \ldots, N$. Placing a suitable grid on the stock price is more difficult, as it is unbounded from above. Let us therefore shrink K to the smaller set $K' = \{(t, x) : t < T,\ x < R\}$ for some $R < \infty$, and we discretize $[0, R]$ into intervals of length $\delta = R/M$, i.e., we work with the stock prices $k\delta$, $k = 0, \ldots, M$. If we choose R sufficiently large, then we expect that the solution of the optimal stopping problem in the set K' will be close to the solution in the set K. In the following we will thus consider the optimal stopping problem in K'.

We now proceed by introducing finite differences. Let us set

$$\frac{\partial V(t, x)}{\partial t} \mapsto \frac{V_{\delta,\Delta}(t, x) - V_{\delta,\Delta}(t - \Delta, x)}{\Delta}$$

for the time derivative,

$$\frac{\partial V(t, x)}{\partial x} \mapsto \frac{V_{\delta,\Delta}(t, x + \delta) - V_{\delta,\Delta}(t, x)}{\delta}$$

for the spatial derivative, and we approximate the second derivative by

$$\frac{\partial^2 V(t,x)}{\partial x^2} \mapsto \frac{V_{\delta,\Delta}(t,x+\delta) - 2\,V_{\delta,\Delta}(t,x) + V_{\delta,\Delta}(t,x-\delta)}{\delta^2}.$$

Now note that the variational inequality can equivalently be written as

$$\min\left\{\left(\frac{\partial}{\partial t} + \frac{\sigma^2 x^2}{2}\frac{\partial^2}{\partial x^2} + \mu x\,\frac{\partial}{\partial x} - r\right)V(t,x),\ \frac{-(K-x)^+ - V(t,x)}{\Delta}\right\} = 0$$

(convince yourself that this is true!), and we can shift time by Δ in the second term without modifying the $\Delta \to 0$ limit (we will shortly see why this is desirable). Substituting into the variational inequality and rearranging gives

$$\begin{aligned} V_{\delta,\Delta}(t-\Delta,x) = \min\Bigg\{-(K-x)^+,\ &\left(1 - \frac{\Delta\sigma^2 x^2}{\delta^2} - \frac{\Delta\mu x}{\delta} - \Delta r\right)V_{\delta,\Delta}(t,x) \\ &+ \left(\frac{\Delta\sigma^2 x^2}{2\delta^2} + \frac{\Delta\mu x}{\delta}\right)V_{\delta,\Delta}(t,x+\delta) + \frac{\Delta\sigma^2 x^2}{2\delta^2}\,V_{\delta,\Delta}(t,x-\delta)\Bigg\}. \end{aligned}$$

Note that this is a backwards in time recursion for $V_{\delta,\Delta}(t,x)$! (It is for this reason that we shifted the terminal cost term in the variational inequality by Δ: if we did not do this, the right hand side would depend on $V_{\delta,\Delta}(t-\Delta,x)$). It remains to specify boundary conditions, but this follows directly from proposition 8.1.8: we should set $V_{\delta,\Delta}(t,x) = -(K-x)^+$ on the boundary of K', i.e., whenever $t = T$ or $x = R$.

We now claim that this discretized equation is itself the dynamic programming equation for an optimal stopping problem for a discrete time Markov chain on a finite state space, provided that Δ is sufficiently small that $1 - \Delta\sigma^2 M^2 - \Delta\mu M - \Delta r \geq 0$.

Proposition 8.1.11. *Let x_k, $k = 0,\dots,N$ be a Markov chain on the state space $\{n\delta : n = 0,\dots,M\}$ with the following transition probabilities for $n < M$:*

$$\begin{aligned} \mathbb{P}(x_k = n\delta | x_{k-1} = n\delta) &= \frac{1 - \Delta\sigma^2 n^2 - \Delta\mu n - \Delta r}{1 - \Delta r}, \\ \mathbb{P}(x_k = (n+1)\delta | x_{k-1} = n\delta) &= \frac{\Delta\sigma^2 n^2 + 2\Delta\mu n}{2 - 2\Delta r}, \\ \mathbb{P}(x_k = (n-1)\delta | x_{k-1} = n\delta) &= \frac{\Delta\sigma^2 n^2}{2 - 2\Delta r}, \end{aligned}$$

and all other transition probabilities are zero. For the state $n = M$ (so $n\delta = R$), let $\mathbb{P}(x_k = R | x_{k-1} = R) = 1$ (so the boundary R is an absorbing state). Moreover, let

$$H[\tau] = \mathbb{E}\left[-(1-\Delta r)^\tau (K - x_\tau)^+\right]$$

for any stopping time $\tau \leq N$ for the filtration generated by x_k (so τ is an $\{0,\dots,N\}$-valued random variable). Denote $D = \{(k, n\delta) : V_{\delta,\Delta}(k\Delta, n\delta) + (K - x_\tau)^+ < 0\}$. Then $\tau^* = \inf\{k : (k, x_k) \notin D\}$ is an optimal stopping rule for the cost $H[\tau]$ in the sense that $H[\tau^*] = \mathbb{E}(V_{\delta,\Delta}(0,x_0)) \leq H[\tau]$ for any stopping time $\tau \leq N$.

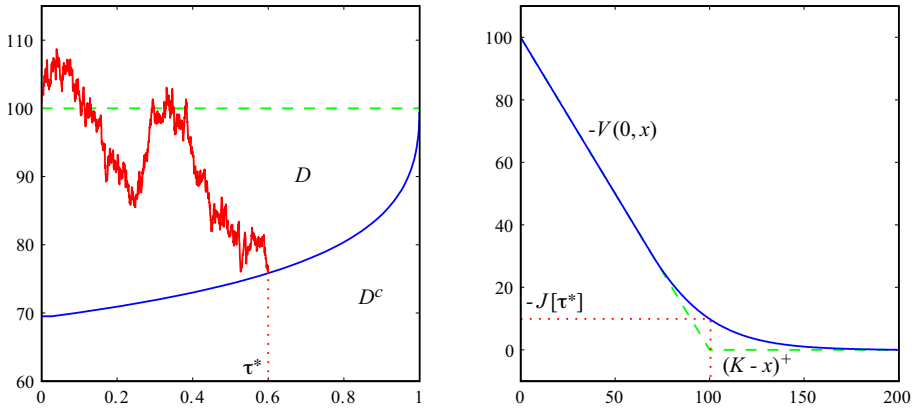

Figure 8.1. Numerical solution of example 8.1.9 with $T = 1$, $K = 100$, $\sigma = .3$, $\mu = r = .05$, and $R = 200$. In the left plot, the boundary ∂D of the continuation region D is plotted in blue, while the contract price K is shown in green (the horizontal axis is time, the vertical axis is price). A single sample path of the stock price, started at $S_0 = 100$, is shown in red; the optimal stopping rule says that we should stop when the stock price hits the curve ∂D. In the right plot, the value function $-V(t, x)$ is plotted in blue for $t = 0$ (the horizontal axis is stock price, the vertical axis is payoff). For an initial stock price of 100 dollars, we see that the option should be priced at approximately ten dollars. Note that the exercise boundary ∂D intersects the line $t = 0$ precisely at the point where $-V(0, x)$ and $(K - x)^+$ (shown in green) diverge.

Proof. Let us write $P_{mn} = \mathbb{P}(x_k = n\delta | x_{k-1} = m\delta)$. Then

$$\begin{aligned}\mathbb{E}(V_{\delta,\Delta}((k-1)\Delta, x_{k-1}) - (1-\Delta r)V_{\delta,\Delta}(k\Delta, x_k)|x_{k-1} = m\delta)\\ = V_{\delta,\Delta}((k-1)\Delta, m\delta) - (1-\Delta r)\sum_n P_{mn}V_{\delta,\Delta}(k\Delta, n\delta) \le 0,\end{aligned}$$

where we have used the equation for $V_{\delta,\Delta}((k-1)\Delta, m\delta)$. In particular, we find that

$$\mathbb{E}(V_{\delta,\Delta}((k-1)\Delta, x_{k-1}) - (1-\Delta r)\, V_{\delta,\Delta}(k\Delta, x_k)|\mathcal{F}_{k-1}) \le 0$$

by the Markov property, where $\mathcal{F}_k = \sigma\{x_\ell : \ell \le k\}$. Now note that $I_{\tau \ge k}$ is $\mathcal{F}_{k-1}$-measurable, as τ is a stopping time. Multiplying by $I_{\tau\ge k}\,(1-\Delta r)^{k-1}$ and taking the expectation gives

$$\mathbb{E}(\{(1-\Delta r)^{k-1}\, V_{\delta,\Delta}((k-1)\Delta, x_{k-1}) - (1-\Delta r)^k\, V_{\delta,\Delta}(k\Delta, x_k)\}\, I_{\tau\ge k}) \le 0.$$

Now sum over k from 1 to N. This gives

$$\mathbb{E}(V_{\delta,\Delta}(0, x_0)) \le \mathbb{E}((1-\Delta r)^\tau\, V_{\delta,\Delta}(\tau\Delta, x_\tau)) \le \mathbb{E}(-(1-\Delta r)^\tau\,(K - x_\tau)^+) = H[\tau],$$

where we have used the equation for $V_{\delta,\Delta}(t, x)$ again. But repeating the same argument with τ^* instead of τ, the inequalities are replaced by equalities and we find that $\mathbb{E}(V_{\delta,\Delta}(0, x_0)) = H[\tau^*]$. Thus τ^* is indeed an optimal stopping time for the discrete problem. □

The numerical solution of the problem is shown in figure 8.1. Evidently the boundary ∂D of the continuation region is a curve, and D is the area above the curve. The

optimal time to exercise the option is the time at which the stock price first hits ∂D (provided that the inital stock price lies above the curve), and we can read off the optimal cost (and hence the fair price of the option) from the value function $V(0,x)$.

8.2 Partial observations: the modification problem

Just as in stochastic control, we do not always have access to the system state X_t in optimal stopping problems. When only a noisy observation process Y_t is available, our only option is to base our stopping decisions on that process. In other words, in this case we wish to minimize the cost $J[\tau]$ over all $\mathcal{F}_t^Y$-stopping times τ (where $\mathcal{F}_t^Y = \sigma\{Y_s : s \le t\}$), rather than over all $\mathcal{F}_t$-stopping times. By definition, this ensures that $\{\omega : \tau(\omega) \le t\} \in \mathcal{F}_t^Y$, so that we can decide when to stop based purely on the observation history. We will encounter examples of such problems in the next two sections, where we discuss applications of optimal stopping in statistics.

We will solve the partial observation problem, as usual, by using the separation principle. However, there is a subtlety in applying this procedure to optimal stopping problems. Recall that we are working with the cost functional

$$J[\tau] = \mathbb{E}\left[\int_0^\tau e^{-\lambda s}\, w(X_s)\, ds + e^{-\lambda\tau}\, z(X_\tau)\right],$$

and assume that τ is an $\mathcal{F}_t^Y$-stopping time. As before, we would like to use the tower property of the conditional expectation to express this cost directly in terms of the filter. Consider first the integral term. If w is nonnegative, for example (so we can apply Fubini's theorem—clearly this can be weakened), we obtain

$$\begin{aligned}
\mathbb{E}\left[\int_0^\tau e^{-\lambda s}\, w(X_s)\, ds\right] &= \mathbb{E}\left[\int_0^\infty I_{s<\tau}\, e^{-\lambda s}\, w(X_s)\, ds\right] \\
&= \int_0^\infty \mathbb{E}(I_{s<\tau}\, e^{-\lambda s}\, w(X_s))\, ds = \int_0^\infty \mathbb{E}(I_{s<\tau}\, e^{-\lambda s}\, \mathbb{E}(w(X_s)|\mathcal{F}_s^Y))\, ds \\
&= \mathbb{E}\left[\int_0^\infty I_{s<\tau}\, e^{-\lambda s}\, \mathbb{E}(w(X_s)|\mathcal{F}_s^Y)\, ds\right] = \mathbb{E}\left[\int_0^\tau e^{-\lambda s}\, \mathbb{E}(w(X_s)|\mathcal{F}_s^Y)\, ds\right],
\end{aligned}$$

where we have used that $I_{s<\tau}$ is $\mathcal{F}_s^Y$-measurable. The second term is more difficult, however. Define $\pi_t(f) = \mathbb{E}(f(X_t)|\mathcal{F}_t^Y)$. Ultimately, we would like to write

$$J[\tau] = \mathbb{E}\left[\int_0^\tau e^{-\lambda s}\, \pi_s(w)\, ds + e^{-\lambda\tau}\, \pi_\tau(z)\right];$$

if this is true, then the partially observed problem reduces to a completely observed optimal stopping problem for the filter. However, it is not at all obvious that

$$\mathbb{E}(e^{-\lambda\tau}\, z(X_\tau)) = \mathbb{E}(e^{-\lambda\tau}\, \pi_\tau(z)) = \mathbb{E}(e^{-\lambda\tau}\, \mathbb{E}(z(X_t)|\mathcal{F}_t^Y)|_{t=\tau}).$$

In fact, at this point, this expression is neither true or false—it is meaningless!

To understand this point, let us revisit the definition of the conditional expectation. Recall that for any random variable $X \in \mathcal{L}^1$ and σ-algebra $\mathcal{F}$, the conditional expectation $\mathbb{E}(X|\mathcal{F})$ is defined uniquely up to a.s. equivalence. In particular, there may

well be two different random variables A and B, both of which satisfy the definition of $\mathbb{E}(X|\mathcal{F})$; however, in this case, we are guaranteed that $A = B$ a.s.

Now consider the *stochastic process* $t \mapsto \pi_t(f)$. For every time t separately, $\pi_t(f)$ is defined uniquely up to a.s. equivalence. But this means that the process $\pi_t(f)$ is only defined uniquely *up to modification*; in particular, two processes A_t and B_t may both satisfy the definition of $\pi_t(f)$, but nonetheless $\mathbb{P}(A_t = B_t\ \forall\, t) < 1$ (after all, we are only guaranteed that $\mathbb{P}(A_t = B_t) = 1$ for all $t \in [0,\infty[$, and $[0,\infty[$ is an uncountable set). In this case, there may well be a stopping time τ such that $A_\tau \neq B_\tau$: *modification need not preserve the value of a process at stopping times*. See section 2.4 for a discussion on this point. Unfortunately, this means that $\mathbb{E}(z(X_t)|\mathcal{F}_t^Y)|_{t=\tau}$ is a meaningless quantity; it may take very different values, even with nonzero probability, depending on how we choose to define our conditional expectations.

Does this mean that all is lost? Not in the least; it only means that we need to do a little more work in defining the process $\pi_t(f)$. As part of the definition of that process, we will select a *particular* modification which has the following special property: $\pi_\tau(f) = \mathbb{E}(f(X_\tau)|\mathcal{F}_\tau^Y)$ for all $\mathcal{F}_t^Y$-stopping times τ (with $\tau < \infty$ a.s.). The process $\pi_t(f)$ is then no longer "just" the conditional expectation process; this particular modification of the conditional expectation process is known as the *optional projection* of the process $f(X_t)$ onto the filtration $\mathcal{F}_t^Y$. Provided that we work with this particular modification, we can complete the separation argument. After all,

$$\mathbb{E}(e^{-\lambda\tau}\, z(X_\tau)) = \mathbb{E}(e^{-\lambda\tau}\, \mathbb{E}(z(X_\tau)|\mathcal{F}_\tau^Y)) = \mathbb{E}(e^{-\lambda\tau}\, \pi_\tau(z)),$$

where we have used that τ is $\mathcal{F}_\tau^Y$-measurable. Hence the problem is now finally reduced to an optimal stopping problem for the filter—that is, if the filter does indeed compute the optional projection $\pi_t(z)$ (which, as we will see, is the case).

Remark 8.2.1. A general theory of optional projections is developed in detail in Dellacherie and Meyer [DM82, section VI.2] (a brief outline can be found in [RW00b, section VI.7]). We will have no need for this general theory, however; instead, we will follow a simple argument due to Rao [Rao72], which provides everything we need.

Let us begin by recalling the definition of the σ-algebra $\mathcal{F}_\tau^Y$ of events up to and including time τ. We have encountered this definition previously: see definition 2.3.16.

Definition 8.2.2. We define $\mathcal{F}_\tau^Y = \{A : A \cap \{\tau \le t\} \in \mathcal{F}_t^Y \text{ for all } t \le \infty\}$ for any $\mathcal{F}_t^Y$-stopping time τ. Then, in particular, τ is $\mathcal{F}_\tau^Y$-measurable.

To demonstrate where we want to be going, consider our usual observation model

$$dY_t = h(X_t)\, dt + K\, dB_t.$$

In the previous chapter, we found filtering equations for $\pi_t(f) = \mathbb{E}(f(X_t)|\mathcal{F}_t^Y)$ for several different signal models; in all these filters, $\pi_t(f)$ was expressed as the sum of $\mathbb{E}(f(X_0))$, a time integral, and a stochastic integral with respect to the observations. But recall that we have defined both the time integral and the stochastic integrals to have *continuous sample paths*; thus the $\pi_t(f)$ obtained by solving the filtering equations of the previous chapter is not just any version of the conditional expectation:

it is the *unique* modification of the conditional expectation process that has continuous sample paths (uniqueness follows from lemma 2.4.6). We are going to show that it is precisely this modification that is the optional projection.

Proposition 8.2.3. *Let X_t be a process with right-continuous sample paths, and let f be a bounded continuous function. Suppose there is a stochastic process $\pi_t(f)$ with continuous sample paths such that $\pi_t(f) = \mathbb{E}(f(X_t)|\mathcal{F}_t^Y)$ for every t. Then $\pi_\tau(f) = \mathbb{E}(f(X_\tau)|\mathcal{F}_\tau^Y)$ for all $\mathcal{F}_t^Y$-stopping times $\tau < \infty$.*

Remark 8.2.4. The "suppose" part of this result is superfluous: it can be shown that a continuous modification of $\mathbb{E}(f(X_t)|\mathcal{F}_t^Y)$ always exists in this setting [Rao72]. We will not need to prove this, however, as we have already explicitly found a continuous modification of the conditional expectation process, viz. the one given by the filtering equations, in all cases in which we are interested.

To prove this statement, we will begin by proving it for the special case that τ only takes a countable number of values. In this case, the result is independent of modification: after all, the problem essentially reduces to discrete time, where two modifications are always indistinguishable (see section 2.4).

Lemma 8.2.5. *Proposition 8.2.3 holds for any modification of $\pi_t(f)$ whenever τ takes values only in a countable set times $\{t_1, t_2, \ldots\}$.*

Proof. We need to show that $\mathbb{E}(\pi_\tau(f)\, I_B) = \mathbb{E}(f(X_\tau)\, I_B)$ for every $B \in \mathcal{F}_\tau^Y$, and that $\pi_\tau(f)$ is $\mathcal{F}_\tau^Y$-measurable. This establishes the claim by the Kolmogorov definition of the conditional expectation. We begin by demonstrating the first claim. Note that $B = \bigcup_{i\geq 1} B \cap \{\tau = t_i\}$, so

$$\mathbb{E}(\pi_\tau(f)\, I_B) = \sum_{i=1}^{\infty} \mathbb{E}(\pi_\tau(f)\, I_{B\cap\{\tau=t_i\}}), \qquad \mathbb{E}(f(X_\tau)\, I_B) = \sum_{i=1}^{\infty} \mathbb{E}(f(X_\tau)\, I_{B\cap\{\tau=t_i\}}).$$

Hence it suffices to prove $\mathbb{E}(\pi_\tau(f)\, I_{B\cap\{\tau=t_i\}}) = \mathbb{E}(f(X_\tau)\, I_{B\cap\{\tau=t_i\}})$ for every $B \in \mathcal{F}_\tau^Y$ and $i \geq 1$. But by definition, $B_i = B \cap \{\tau = t_i\} \in \mathcal{F}_{t_i}^Y$, so we do indeed find that

$$\mathbb{E}(\pi_\tau(f)\, I_{B_i}) = \mathbb{E}(\pi_{t_i}(f)\, I_{B_i}) = \mathbb{E}(f(X_{t_i})\, I_{B_i}) = \mathbb{E}(f(X_\tau)\, I_{B_i}).$$

To show that $\pi_\tau(f)$ is $\mathcal{F}_\tau^Y$-measurable, note that

$$\{\pi_\tau(f) \in A\} = \bigcup_{i\geq 1} \{\pi_\tau(f) \in A \text{ and } \tau = t_i\} = \bigcup_{i\geq 1} \{\pi_{t_i}(f) \in A\} \cap \{\tau = t_i\}$$

for every Borel set A; hence $\{\pi_\tau(f) \in A\} \cap \{\tau = t_j\} \in \mathcal{F}_{t_j}^Y \subset \mathcal{F}_{t_i}^Y$ for every $j \leq i$, so it follows easily that $\{\pi_\tau(f) \in A\} \in \mathcal{F}_\tau^Y$ (take the union over $j \leq i$). We are done. □

To prove proposition 8.2.3 in its full glory, we can now proceed as follows. Even though τ does not take a countable number of values, we can always approximate it by a sequence of stopping times τ_n such that every τ_n takes a countable number of values and $\tau_n \searrow \tau$. We can now take limits in the previous lemma, and this is precisely where the various continuity assumptions will come in. Before we complete the proof, we need a an additional lemma which helps us take the appropriate limits.

Lemma 8.2.6 (A version of Hunt's lemma). *Let X_n be a sequence of random variables such that $X_n \to X_\infty$ in $\mathcal{L}^2$, let $\mathcal{F}_n$ be a reverse filtration $\mathcal{F}_{n-1} \supset \mathcal{F}_n \supset \mathcal{F}_{n+1} \supset \cdots$, and denote by $\mathcal{F}_\infty = \bigcap_n \mathcal{F}_n$. Then $\mathbb{E}(X_n|\mathcal{F}_n) \to \mathbb{E}(X_\infty|\mathcal{F}_\infty)$ in $\mathcal{L}^2$.*

Proof. Note that we can write $\|\mathbb{E}(X_n|\mathcal{F}_n) - \mathbb{E}(X_\infty|\mathcal{F}_\infty)\|_2 \leq \|\mathbb{E}(X_n - X_\infty|\mathcal{F}_n)\|_2 + \|\mathbb{E}(X_\infty|\mathcal{F}_n) - \mathbb{E}(X_\infty|\mathcal{F}_\infty)\|_2$. But for the first term, we obtain using Jensen's inequality $\|\mathbb{E}(X_n - X_\infty|\mathcal{F}_n)\|_2 \leq \|X_n - X_\infty\|_2 \to 0$ by assumption. Hence it remains to prove that the second term converges to zero. To this end, let us write $F_n = \mathbb{E}(X_\infty|\mathcal{F}_n)$. Then for $m < n$

$$\begin{aligned}\mathbb{E}((F_m - F_n)^2) &= \mathbb{E}(F_m^2) + \mathbb{E}(F_n^2) - 2\,\mathbb{E}(F_n F_m) \\ &= \mathbb{E}(F_m^2) + \mathbb{E}(F_n^2) - 2\,\mathbb{E}(F_n\,\mathbb{E}(F_m|\mathcal{F}_n)) = \mathbb{E}(F_m^2) - \mathbb{E}(F_n^2).\end{aligned}$$

But then we find, in particular, that

$$\sum_{k=m+1}^{n} \mathbb{E}((F_{k-1} - F_k)^2) = \mathbb{E}(F_m^2) - \mathbb{E}(F_n^2) \leq 2\,\mathbb{E}(X_\infty^2) < \infty,$$

so we conclude (let $n \to \infty$, then $m \to \infty$) that F_n is a Cauchy sequence in $\mathcal{L}^2$. But then F_n must converge in $\mathcal{L}^2$ to some random variable F_∞, and it remains to verify that $F_\infty = \mathbb{E}(X_\infty|\mathcal{F}_\infty)$. To this end, note that for every $A \in \mathcal{F}_\infty \subset \mathcal{F}_n$, we have

$$\mathbb{E}(F_\infty\, I_A) = \lim_{n\to\infty} \mathbb{E}(F_n\, I_A) = \lim_{n\to\infty} \mathbb{E}(X_\infty\, I_A) = \mathbb{E}(X_\infty\, I_A)$$

by dominated convergence. On the other hand, $\{F_n\}_{n\geq m}$ is a sequence of $\mathcal{F}_m$-measurable random variables, so the limit (in $\mathcal{L}^2$) of this sequence is also $\mathcal{F}_m$-measurable. But F_∞ is the limit of every such sequence, so F_∞ is $\mathcal{F}_m$ measurable for every m, i.e., F_∞ is $\mathcal{F}_\infty$-measurable, and this establishes the claim by the Kolmogorov definition of the conditional expectation. □

We can finally proceed to the proof of proposition 8.2.3.

Proof of proposition 8.2.3. Define the stopping times $\tau_n = 2^{-n}(\lfloor 2^n \tau \rfloor + 1)$. Then $\tau_n \searrow \tau$, and each τ_n takes a countable number of values. By lemma 8.2.5, we find that $\pi_{\tau_n}(f) = \mathbb{E}(f(X_{\tau_n})|\mathcal{F}^Y_{\tau_n})$ for every n. We would like to take the limit of this expression as $n \to \infty$. The left-hand side is easy: $\pi_{\tau_n}(f) \to \pi_\tau(f)$ by the continuity of the sample paths of $\pi_t(f)$ (which we have assumed). It remains to tackle the right-hand side.

First, we claim that $\mathcal{F}^Y_{\tau_n} \supset \mathcal{F}^Y_{\tau_{n+1}}$. To see this, let $A \in \mathcal{F}^Y_{\tau_{n+1}}$ be arbitrary. Then

$$A \cap \{\tau_n \leq t\} = A \cap \{\tau_{n+1} \leq t\} \cap \{\tau_n \leq t\} \in \mathcal{F}^Y_t,$$

where we have used that $\tau_{n+1} \leq \tau_n$, the definition of $\mathcal{F}^Y_{\tau_{n+1}}$ and that τ_n is an $\mathcal{F}^Y_t$-stopping time. But this holds for every t, so the claim follows by the definition of $\mathcal{F}^Y_{\tau_n}$. We can thus conclude by lemma 8.2.6, the boundedness of f and the right-continuity of $f(X_t)$, that

$$\mathbb{E}(f(X_{\tau_n})|\mathcal{F}^Y_{\tau_n}) \xrightarrow{n\to\infty} \mathbb{E}(f(X_\tau)|\mathcal{G}) \quad \text{in } \mathcal{L}^2, \qquad \mathcal{G} = \bigcap_{n\geq 1} \mathcal{F}^Y_{\tau_n}.$$

We now have to show that $\mathbb{E}(f(X_\tau)|\mathcal{G}) = \mathbb{E}(f(X_\tau)|\mathcal{F}^Y_\tau)$ a.s. Clearly $\mathcal{F}^Y_\tau \subset \mathcal{G}$ (as $\tau < \tau_n$ for all n), so it suffices to show that $\mathbb{E}(\mathbb{E}(f(X_\tau)|\mathcal{G})|\mathcal{F}^Y_\tau) = \mathbb{E}(f(X_\tau)|\mathcal{G})$ a.s. But we know that $\mathbb{E}(f(X_\tau)|\mathcal{G}) = \pi_\tau(f)$ a.s. Hence we are done if we can show that $\pi_\tau(f)$ is $\mathcal{F}^Y_\tau$-measurable.

To see that this is the case, define the stopping times $\sigma_n = \tau_n - 2^{-n}$. Then $\sigma_n \leq \tau$, and $\sigma_n \to \tau$ as $n \to \infty$. But $\pi_{\sigma_n}(f)$ is $\mathcal{F}^Y_{\sigma_n}$-measurable (see the proof of lemma 8.2.5), so it is $\mathcal{F}^Y_\tau$-measurable for every n (as $\sigma_n \leq \tau$ implies $\mathcal{F}^Y_{\sigma_n} \subset \mathcal{F}^Y_\tau$). But then $\pi_\tau(f) = \lim_{n\to\infty} \pi_{\sigma_n}(f)$ (by the continuity of the sample paths) must be $\mathcal{F}^Y_\tau$-measurable. □

8.3 Changepoint detection

We are now ready to treat a statistical application: the detection of a sudden change in white noise. This is known as a *changepoint detection* problem.

The problem is as follows. Our system—and industrial process, a stock, a computer network, etc.—has a parameter which undergoes a sudden change at some random time τ (in the above examples, e.g., a machine breakdown, a stock crash, a denial-of-service attack, etc.). We will assume that τ is exponentially distributed, i.e., $\mathbb{P}(\tau = 0) = \pi_0$ and $\mathbb{P}(\tau > t | \tau > 0) = e^{-\lambda t}$. In an ideal world, we would intervene as soon as the sudden change occurs, i.e., we would like to take some action to correct for the change. Unfortunately, we can not actually see when the change happens; all that is available to us is the noisy observation process

$$dY_t = \gamma\, I_{\tau \le t}\, dt + \sigma\, dB_t, \qquad \mathcal{F}_t^Y = \sigma\{Y_s : s \le t\},$$

where B_t is a Wiener process independent of τ. Our goal is thus to find an $\mathcal{F}_t^Y$-stopping time ϑ, i.e., a stopping time that is decided purely on the basis of the observations, that is as close as possible to the changepoint τ in a suitable sense.

In deciding on a stopping strategy ϑ, however, we have two competing goals. On the one hand, we would like to intervene as soon as possible after the changepoint τ occurs, i.e., we would like to minimize the *expected delay* $\mathbb{E}((\vartheta - \tau)^+)$. On the other hand, it is bad if we decide to intervene before the change has actually occured, i.e., we would like to minimize the *probability of false alarm* $\mathbb{P}(\vartheta < \tau)$. As you can imagine, these goals are in some sense mutually exclusive: if we do not care about false alarms then it is always best to stop at $\vartheta = 0$ (as the delay time is then zero!), while if we do not care about delay then we should intervene at $\vartheta = \infty$ ($\tau < \infty$ a.s., so if we wait long enough we are sure that there will be no false alarm). Generally, however, there is a tradeoff between the two, and it is in this case that the problem becomes nontrivial. To quantify the tradeoff, let us introduce the cost functional

$$J[\vartheta] = \mathbb{P}(\vartheta < \tau) + c\,\mathbb{E}((\vartheta - \tau)^+).$$

The constant $c > 0$ allows us to select the relative merit of minimizing the false alarm probability or the delay time. Our goal is to find an $\mathcal{F}_t^Y$-stopping rule ϑ^* which minimizes $J[\vartheta]$. Using the Shiryaev-Wonham filter, we can turn this into an ordinary optimal stopping problem to which proposition 8.1.3 can be applied.

Remark 8.3.1. There are various variations on the changepoint detection problem, some of which we will discuss at the end of this section. For the time being, however, let us concentrate on solving the problem in this basic form.

Define $\pi_t = \mathbb{P}(\tau \le t | \mathcal{F}_t^Y)$, and recall that the Shiryaev-Wonham filter gives

$$d\pi_t = \frac{\gamma}{\sigma}\,\pi_t(1-\pi_t)\,d\bar{B}_t + \lambda(1-\pi_t)\,dt, \qquad d\bar{B}_t = \sigma^{-1}(dY_t - \gamma\pi_t\,dt).$$

If we write $I_{\tau \le t} = X_t$, then we obtain the more familiar cost

$$J[\vartheta] = \mathbb{E}\left[c\int_0^\vartheta X_s\,ds + (1 - X_\vartheta)\right] = \mathbb{E}\left[\int_0^\vartheta c\,\pi_s\,ds + (1-\pi_\vartheta)\right],$$

where we have used the result of the previous section (note that $I_{\tau \leq t}$ does not have continuous sample paths, but it does have right-continuous sample paths). We can thus apply proposition 8.1.3 with $K = [0, 1]$, and the variational inequality reads

$$\min\left\{\frac{\gamma^2 x^2(1-x)^2}{2\sigma^2}\frac{\partial^2 V(x)}{\partial x^2} + \lambda(1-x)\frac{\partial V(x)}{\partial x} + cx,\ 1 - x - V(x)\right\} = 0.$$

Perhaps remarkably, this problem has a (somewhat) explicit solution.

To begin, recall that once we have obtaind a suitable solution $V(x)$ to this problem, the interval $[0, 1]$ is divided into the continuation region $D = \{x : V(x) < 1 - x\}$ and the stopping region $D^c = \{x : V(x) = 1 - x\}$. On the former, we must have $\mathscr{L}V(x)+cx = 0$, while on the latter we must have $\mathscr{L}V(x)+cx \geq 0$. In particular, we can use the latter requirement to find a necessary condition on the set D: substituting $V(x) = 1 - x$ into the inequality, we find that it must be the case that $x \geq \lambda/(c + \lambda)$ for any $x \in D^c$. In particular, this implies that $[0, \lambda/(c + \lambda)[\subset D$.

Let us now try to solve for $V(x)$ on D. Note that $\mathcal{L}V(x) + cx = 0$ gives

$$\frac{\partial U(x)}{\partial x} = -\frac{2\sigma^2}{\gamma^2}\left(\frac{\lambda}{x^2(1-x)}U(x) + \frac{c}{x(1-x)^2}\right) \quad (x > 0), \qquad U(0) = 0,$$

where $U(x) = \partial V(x)/\partial x$. This is an (admittedly nasty) ordinary differential equation, which does however have locally Lipschitz coefficients on $]0, 1[$; if we require[1] $U(x) \to 0$ as $x \to 0$ (so that $V(x)$ is C^1 at $x = 0$), we obtain the unique solution

$$U(x) = -\frac{2\sigma^2 c}{\gamma^2}\, e^{-\frac{2\sigma^2\lambda}{\gamma^2}(\log(\frac{x}{1-x})-\frac{1}{x})} \int_0^x \frac{e^{\frac{2\sigma^2\lambda}{\gamma^2}(\log(\frac{y}{1-y})-\frac{1}{y})}}{y(1-y)^2}\, dy.$$

Let us verify some elementary properties of this equation.

Lemma 8.3.2. *The function $U(x)$, as defined above, has the following properties:*

1. *$U(x) \leq 0$;*
2. *$U(x)$ is C^1 on $[0, 1[$;*
3. *$U(x) \to 0$ as $x \searrow 0$; and*
4. *$U(x) \to -\infty$ as $x \nearrow 1$.*

Proof. That $U(x) \leq 0$ is trivial. It is easily verified that the integrand of the integral in the expression for $U(x)$ is bounded on every set of the form $[0, x]$ with $x < 1$. Thus $U(x)$ is clearly well-defined for every $x \in]0, 1[$, and it follows directly that $U(x)$ is C^1 in $]0, 1[$. The behavior of $U(x)$ as $x \searrow 0$ or $x \nearrow 1$ follows by an application of l'Hospital's rule (with the integral in the numerator and the exponential prefactor in the denominator). It remains to show that $U(x)$ is differentiable at $x = 0$ (so $U(x)$ is C^1 on $[0, 1[$); this follows by applying l'Hospital's rule to $U(x)/x$ (with the integral in the numerator and the prefactor in the denominator). □

Here is another very useful property of $U(x)$, which is not immediately obvious.

[1] We have to require this to apply proposition 8.1.3, as we know that $0 \in D$ and $V(x)$ must be C^1.

Lemma 8.3.3. *$U(x)$ is strictly decreasing.*

Proof. Using the differential equation satisfied by $U(x)$, it follows that

$$\frac{\partial U(x)}{\partial x} < 0 \qquad \Longleftrightarrow \qquad -U(x) < \frac{c}{\lambda}\frac{x}{1-x}.$$

We would like to show that this is in fact the case. Using the expression for $U(x)$, this becomes

$$\frac{2\sigma^2\lambda}{\gamma^2}\, e^{-\frac{2\sigma^2\lambda}{\gamma^2}(\log(\frac{x}{1-x})-\frac{1}{x})} \int_0^x \frac{e^{\frac{2\sigma^2\lambda}{\gamma^2}(\log(\frac{y}{1-y})-\frac{1}{y})}}{y(1-y)^2}\, dy < \frac{x}{1-x}.$$

The trick is to note that we have the identity

$$\frac{d}{dy}\, e^{\frac{2\sigma^2\lambda}{\gamma^2}(\log(\frac{y}{1-y})-\frac{1}{y})} = \frac{2\sigma^2\lambda}{\gamma^2}\frac{e^{\frac{2\sigma^2\lambda}{\gamma^2}(\log(\frac{y}{1-y})-\frac{1}{y})}}{y^2(1-y)},$$

so that it evidently remains to prove that

$$e^{-\frac{2\sigma^2\lambda}{\gamma^2}(\log(\frac{x}{1-x})-\frac{1}{x})} \int_0^x \frac{y}{1-y}\,\frac{d}{dy}\, e^{\frac{2\sigma^2\lambda}{\gamma^2}(\log(\frac{y}{1-y})-\frac{1}{y})}\, dy < \frac{x}{1-x}.$$

But this is clearly true for all $0 < x < 1$, and thus the proof is complete. □

We can now finally complete the solution of the optimal stopping problem. We need to use the principle of smooth fit to determine D. Note that for x on the boundary of D, we must have $U(x) = -1$ in order to make $V(x)$ be C^1. But the previous lemmas demonstrate that there is a *unique* point $\pi^* \in]0,1[$ such that $U(\pi^*) = -1$: there is at least one such point (as $U(0) \geq -1 \geq U(1)$), and uniqueness follows as $U(x)$ is decreasing. We have thus established, in particular, that the continuation region must be of the form $D = [0, \pi^*[$, and the remainder of the argument is routine.

Theorem 8.3.4 (Changepoint detection). *Let $\pi^* \in]0,1[$ be the unique point such that $U(\pi^*) = -1$, and define the concave function $V(x)$ as*

$$V(x) = \begin{cases} 1 - \pi^* + \int_{\pi^*}^x U(y)\, dy & \text{for } x \in [0, \pi^*[, \\ 1 - x & \text{for } x \in [\pi^*, 1]. \end{cases}$$

Then $V(x)$ is C^1 on $[0,1]$, C^2 on $[0,1]\backslash\{\pi^*\}$, and

$$\min\left\{ \frac{\gamma^2 x^2 (1-x)^2}{2\sigma^2}\frac{\partial^2 V(x)}{\partial x^2} + \lambda(1-x)\frac{\partial V(x)}{\partial x} + cx,\ 1 - x - V(x) \right\} = 0.$$

In particular, the stopping rule $\vartheta^* = \inf\{t : \pi_t \notin [0, \pi^*[\}$ is optimal in that it minimizes the cost $J[\vartheta]$ in the class of $\mathcal{F}_t^Y$-stopping times.

Proof. The various smoothness properties of $V(x)$ follow from the lemmas proved above. That $V(x)$ is concave follows as its second derivative is nonpositive. Next, let us show that (as expected) $\pi^* > \lambda/(c+\lambda)$. To this end, it suffices to substitute $U(x) = -1$ into the equation for $\partial U(x)/\partial x$, and to use the fact that the latter is negative. But then it follows directly that the variational inequality is satisfied on $[\pi^*, 1]$. That $\mathscr{L}V(x) + cx = 0$ on $[0, \pi^*[$ follows from the

definition of $U(x)$, while that $V(x) < 1 - x$ on $[0, \pi^*]$ follows from the concavity of $V(x)$. Hence we have established that the variational inequality is satisfied on $[0, 1]$.

We now invoke proposition 8.1.3 with $K = [0, 1]$. Clearly $V(x)$ is sufficiently smooth, and as $V(x)$ and both its derivatives are bounded, it remains by lemma 6.3.4 to show that $\mathbb{E}(\vartheta^*) < \infty$. To this end, define $\alpha_t = -\log(1 - \pi_t)$. Then Itô's rule gives

$$d\alpha_t = \frac{\gamma}{\sigma}\pi_t \, d\bar{B}_t + \lambda \, dt + \frac{\gamma^2}{2\sigma^2}\,\pi_t^2 \, dt.$$

Hence we obtain, in particular,

$$\mathbb{E}(\alpha_{t\wedge\vartheta^*}) \geq \alpha_0 + \lambda\,\mathbb{E}(t \wedge \vartheta^*).$$

But $\alpha_{t\wedge\vartheta^*} \leq -\log(1 - \pi^*)$, so $\mathbb{E}(\vartheta^*) < \infty$ by monotone convergence. We are done. □

We now have a complete solution to the basic changepoint detection problem for our model. The rest of this section discusses some variations on this theme.

Example 8.3.5 (Variational formulation). We have discussed what is known as the "Bayesian" form of the changepoint detection problem: we have quantified the trade-off between false alarm probability and expected delay by minimizing a weighted sum of the two. Sometimes, however, a "variational" form of the problem is more appropriate. The latter asks the following: *in the class of $\mathcal{F}_t^Y$-stopping rules ϑ with a false alarm probability of at most $\alpha \in]0, 1[$, what is the stopping rule that minimizes the expected delay?* With the solution to the Bayesian problem in hand, we can now solve the variational problem using a method similar to the one used in example 6.4.7.

Corollary 8.3.6. *Let $\alpha \in]0, 1[$. Then amongst those $\mathcal{F}_t^Y$-stopping times ϑ such that $\mathbb{P}(\vartheta < \tau) \leq \alpha$, the expected delay is minimized by $\vartheta^* = \inf\{t : \pi_t \notin [0, 1 - \alpha[\}$.*

Proof. First, we claim that we can choose the Bayesian cost $J[\vartheta]$ in such a way that $\pi^* = 1-\alpha$ in the previous theorem, i.e., there exists a constant $c_\alpha > 0$ such that $U(1 - \alpha) = -1$ for $c = c_\alpha$. This is trivially seen, however, as $U(x)$ is directly proportional to c. Denote by $J_\alpha[\vartheta]$ the cost with $c = c_\alpha$; evidently $\vartheta^* = \inf\{t : \pi_t \notin [0, 1 - \alpha[\}$ minimizes $J_\alpha[\vartheta]$.

Next, we claim that $\mathbb{P}(\vartheta^* < \tau) = \alpha$ whenever $\pi_0 < 1 - \alpha$. To see this, note that using proposition 8.2.3, we can write $\mathbb{P}(\vartheta^* < \tau) = 1 - \mathbb{E}(\mathbb{P}(\tau \leq \vartheta^* | \mathcal{F}_{\vartheta^*}^Y)) = 1 - \mathbb{E}(\pi_{\vartheta^*}) = \alpha$. Whenever $\pi_0 \geq 1 - \alpha$, it must be the case that $\vartheta^* = 0$, so we find $\mathbb{P}(\vartheta^* < \tau) = 1 - \pi_0$. For the latter case, however, the result holds trivially. To see this, note that $\vartheta^* = 0$ satisfies $\mathbb{P}(\vartheta < \tau) \leq \alpha$, while the expected delay is zero for this case. As a smaller delay is impossible, $\vartheta^* = 0$ is indeed optimal in the class of stopping rules in which we are interested.

It thus remains to consider the case when $\pi_0 < 1 - \alpha$. To this end, let ϑ be an arbitrary $\mathcal{F}_t^Y$-stopping time with $\mathbb{P}(\vartheta < \tau) \leq \alpha$. Then $J_\alpha[\vartheta^*] \leq J_\alpha[\vartheta]$ gives

$$\begin{aligned}\alpha + c_\alpha\,\mathbb{E}((\vartheta^* - \tau)^+) &= \mathbb{P}(\vartheta^* < \tau) + c_\alpha\,\mathbb{E}((\vartheta^* - \tau)^+) \\ &\leq \mathbb{P}(\vartheta < \tau) + c_\alpha\,\mathbb{E}((\vartheta - \tau)^+) \leq \alpha + c_\alpha\,\mathbb{E}((\vartheta - \tau)^+).\end{aligned}$$

Thus $\mathbb{E}((\vartheta^*-\tau)^+) \leq \mathbb{E}((\vartheta-\tau)^+)$, so ϑ cannot have a smaller delay than ϑ^*. We are done. □

Note that the variational problem is, in some sense, much more intuitive than its Bayesian counterpart: given that we can tolerate a fixed probability of false alarm α, it is best not to stop until the conditional probability of being in error drops below α.

Example 8.3.7 (Expected miss criterion). The cost $J[\vartheta]$ is quite general, and contains seemingly quite different cost functionals as special cases. To demonstrate this point, let us show how to obtain a stopping rule ϑ^* that minimizes the *expected miss criterion* $J'[\vartheta] = \mathbb{E}(|\vartheta - \tau|)$. In the absence of explicit false alarm/expected delay preferences, this is arguably the most natural cost functional to investigate!

The solution of this problem is immediate if we can rewrite the cost $J'[\vartheta]$ in terms of $J[\vartheta]$ (for some suitable c). This is a matter of some clever manipulations, combined with explicit use of the Shiryaev-Wonham filter. We begin by noting that

$$J'[\vartheta] = \mathbb{E}(\tau + \vartheta - 2\,\tau \wedge \vartheta) = \frac{1-\pi_0}{\lambda} + \mathbb{E}\left[\int_0^\vartheta \{1 - 2\,I_{s<\tau}\}\,ds\right],$$

where we have used $\mathbb{E}(\tau) = (1-\pi_0)/\lambda$. Using the the tower property, we obtain

$$J'[\vartheta] = \frac{1-\pi_0}{\lambda} + \mathbb{E}\left[\int_0^\vartheta \{2\,I_{\tau\le s} - 1\}\,ds\right] = \frac{1-\pi_0}{\lambda} + \mathbb{E}\left[\int_0^\vartheta \{\pi_s - (1-\pi_s)\}\,ds\right].$$

Now note that the Shiryaev-Wonham equation gives

$$\pi_\vartheta = \pi_0 + \lambda\int_0^\vartheta (1-\pi_s)\,ds + \frac{\gamma}{\sigma}\int_0^\vartheta \pi_s(1-\pi_s)\,d\bar{B}_s.$$

In particular, if we restrict to stopping times ϑ with $\mathbb{E}(\vartheta) < \infty$ (this is without loss of generality, as it is easy to see that $J'[\vartheta] = \infty$ if $\mathbb{E}(\vartheta) = \infty$), then lemma 6.3.4 gives

$$\lambda\,J'[\vartheta] = 1 - \pi_0 + \mathbb{E}\left[\int_0^\vartheta \lambda\pi_s\,ds + \pi_0 - \pi_\vartheta\right] = J_\lambda[\vartheta],$$

where $J_\lambda[\vartheta]$ is our usual cost $J[\vartheta]$ with $c = \lambda$. Evidently $J'[\vartheta]$ and $J_\lambda[\vartheta]$ differ only by a constant factor, and hence their minima are the same. It thus remains to invoke theorem 8.3.4 with $c = \lambda$, and we find that $\vartheta^* = \inf\{t : \pi_t \notin [0, \pi^*[\}$ for suitable π^*.

Let us complete this section with an interesting application from [RH06].

Example 8.3.8 (Optimal stock selling). The problem which we wish to consider is how to best make money off a "bubble stock". Suppose we own a certain amount of stock in a company that is doing well—the stock price increases on average. However, at some random point in time τ the company gets into trouble (the "bubble bursts"), and the stock price starts falling rapidly from that point onward. Concretely, you can imagine a situation similar to the dot-com bubble burst in early 2000.

It seems evident that we should sell our stock before the price has dropped too far, otherwise we will lose a lot of money. However, all we can see is the stock price: if the stock price starts dropping, we are not sure whether it is because the bubble burst or whether it is just a local fluctuation in the market (in which case the stock price will go up again very shortly). The problem is thus to try to determine, based only on the observed stock prices, when is the best time to sell our stock.

Let us introduce a simple model in which we can study this problem. The total amount of money we own in stock is denoted S_t, and satisfies

$$dS_t = \mu_t S_t\,dt + \sigma S_t\,dW_t.$$

Prior to the burst time τ, the stock is making money: we set $\mu_t = a > 0$ for $t < \tau$. After the burst, the stock loses money: we set $\mu_t = -b < 0$ for $\tau \le t$. In particular,

$$dS_t = (a\,I_{t<\tau} - b\,I_{\tau\le t})S_t\,dt + \sigma S_t\,dW_t.$$

Denote by $\mathcal{F}_t^S = \sigma\{S_s : s \le t\}$ the filtration generated by the stock price, and we choose τ to be an exponentially distributed random variable, i.e., $\mathbb{P}(\tau = 0) = \pi_0$ and $\mathbb{P}(\tau > t|\tau > 0) = e^{-\lambda t}$, which is independent of the Wiener process W_t. Our goal is to maximize the expected utility $\mathbb{E}(u(S_\vartheta))$ from selling at time ϑ, i.e., we seek to minimize the cost $J'[\vartheta] = \mathbb{E}(-u(S_\vartheta))$ in the class of $\mathcal{F}_t^S$-stopping rules (see example 6.2.4 for a discussion of utility). For simplicity we will concentrate here on the Kelly criterion $u(x) = \log(x)$ (the risk-neutral case can be treated as well, see [RH06]).

We begin by rewriting the cost in a more convenient form. We will restrict ourselves throughout to stopping times with $\mathbb{E}(\vartheta) < \infty$ (and we seek an optimal stopping rule in this class), so that we can apply lemma 6.3.4. Using Itô's rule, we then obtain

$$J'[\vartheta] = \mathbb{E}\left[\int_0^\vartheta \left\{\frac{\sigma^2}{2} - a\,I_{s<\tau} + b\,I_{\tau\le s}\right\} ds\right] - \log(S_0),$$

where we presume that S_0 (our starting capital) is non-random. It follows that

$$J'[\vartheta] = \mathbb{E}\left[\int_0^\vartheta \left\{\frac{\sigma^2}{2} - a + (a+b)\,\mathbb{P}(\tau \le s|\mathcal{F}_s^S)\right\} ds\right] - \log(S_0),$$

where we have used the tower property of the conditional expectation.

Let us now introduce the process $Y_t = \log(S_t) - \log(S_0) + (\frac{1}{2}\sigma^2 - a)\,t$. Then

$$dY_t = -(a+b)\,I_{\tau\le t}\,dt + \sigma\,dW_t = \gamma\,I_{\tau\le t}\,dt + \sigma\,dW_t, \qquad \gamma = -(a+b).$$

But clearly we can transform back and forth between S_t and Y_t without losing any information, so in particular $\mathcal{F}_t^S = \mathcal{F}_t^Y = \sigma\{Y_s : s \le t\}$, and we can thus conclude that $\mathbb{P}(\tau \le t|\mathcal{F}_t^S) = \mathbb{P}(\tau \le t|\mathcal{F}_t^Y) = \pi_t$ satisfies the Shiryaev-Wonham equation:

$$d\pi_t = \frac{\gamma}{\sigma}\,\pi_t(1-\pi_t)\,d\bar{B}_t + \lambda(1-\pi_t)\,dt, \qquad d\bar{B}_t = \sigma^{-1}(dY_t - \gamma\pi_t\,dt).$$

We can now transform the cost $J'[\vartheta]$ into the cost $J[\vartheta]$ of the changepoint detection problem as in the previous example. To this end, we rewrite $J'[\vartheta]$ suggestively as

$$J'[\vartheta] = \mathbb{E}\left[\int_0^\vartheta \left\{\left(\frac{\sigma^2}{2} - a\right)(1-\pi_s) + \left(\frac{\sigma^2}{2} + b\right)\pi_s\right\} ds\right] - \log(S_0).$$

Using the Shiryaev-Wonham equation, we obtain

$$J'[\vartheta] = \mathbb{E}\left[\frac{1}{\lambda}\left(a - \frac{\sigma^2}{2}\right)(\pi_0 - \pi_\vartheta) + \int_0^\vartheta \left(\frac{\sigma^2}{2} + b\right)\pi_s\,ds\right] - \log(S_0).$$

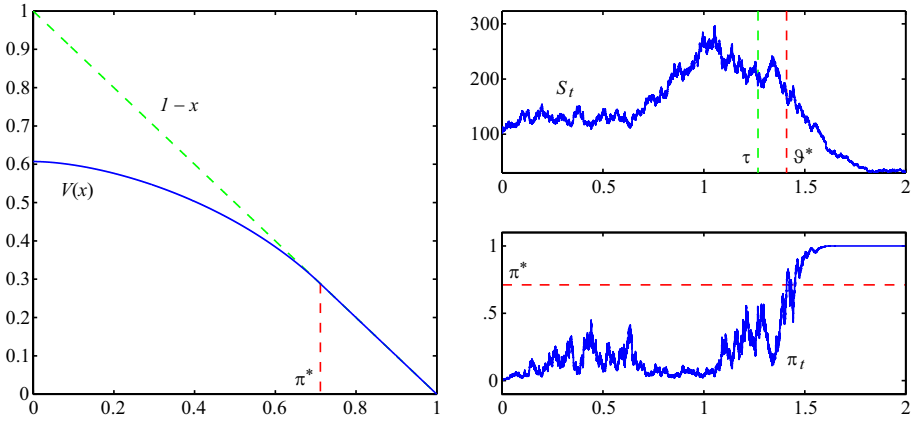

Figure 8.2. Solution of example 8.3.8 with $a = 1$, $b = 2$, $\sigma = .6$, $S_0 = 100$, $\lambda = 1$, and $\pi_0 = 0$. The left plot shows $V(x)$ of theorem 8.3.4; note that the threshold π^* is precisely the point where $V(x)$ and $1 - x$ diverge. A simulation on the time interval $[0, 2]$ of the stock price S_t (which starts tanking at time τ) and the filter π_t, with the optimal stopping strategy in action, is shown on the right. The optimal time to sell is the first time the filter exceeds π^*.

Things are starting to look up—this is almost the changepoint detection problem!

To proceed, we need to distinguish between two cases. The first case is when $2a \leq \sigma^2$. In this case, the problem becomes essentially trivial; indeed, you can read of from the expression for $J'[\vartheta]$ above that the optimal stopping rule for this case tries to simultaneously minimize the expected delay, and maximize the probability of false alarm. This is easily accomplished by setting $\vartheta^* = 0$, and this is indeed the optimal stopping rule when $2a \leq \sigma^2$. It thus remains to consider the nontrivial case $2a > \sigma^2$.

Define the constants $q = (2a - \sigma^2)/2\lambda$ and $c = (\sigma^2 + 2b)/2q$, and note in particular that $q, c > 0$ when $2a > \sigma^2$ (which we now assume). Then we can write

$$q^{-1} J'[\vartheta] = \mathbb{E}\left[1 - \pi_\vartheta + \int_0^\vartheta c\pi_s \, ds\right] - q^{-1} \log(S_0) + \pi_0 - 1.$$

In particular, we find that $J'[\vartheta] = q\, J[\vartheta] - \log(S_0) - q\,(1 - \pi_0)$, where $J[\vartheta]$ is our usual changepoint detection cost (with $c = (\sigma^2 + 2b)/2q$). But as $q > 0$, it is clear that the $\mathcal{F}_t^S$-stopping rule ϑ^* that minimizes $J'[\vartheta]$ coincides with the $\mathcal{F}_t^Y$-stopping rule that minimizes $J[\vartheta]$. Hence the optimal time ϑ^* to sell our stock can be found directly by substituting the appropriate values into theorem 8.3.4.

The function $V(x)$ of theorem 8.3.4 is shown in figure 8.2 for a particular set of parameter values for the stock selling problem. Note that the value function does indeed have the desired behavior, and we can read off the boundary π^* of the continuation region. A numerical simulation shows the stock selling problem in action.

8.4 Hypothesis testing

Another classic optimal stopping problem from statistics is the *sequential testing of hypotheses*. We will develop this theory here in the simplest continuous time setting.

The model in this case is very simple. Suppose that we are sent a single bit through a noisy channel. The observation process is then given by

$$dY_t = \gamma X\, dt + \sigma\, dB_t, \qquad \mathcal{F}_t^Y = \sigma\{Y_s : s \leq t\},$$

where X (the bit) is either zero or one. We would like to determine the value of X on the basis of the observations, i.e., we would like to accept one of the two *hypotheses* $X = 0$ or $X = 1$, and we would like to do this in such a way that the probabilities of selecting the wrong hypothesis (we accept the hypothesis $X = 0$ when in fact $X = 1$, and vice versa) are as small as possible. On a fixed time horizon $[0, T]$, it is well known how to do this: the *Neyman-Pearson test* characterizes this case completely.

The problem becomes more interesting, however, when we do not fix the observation interval $[0, T]$, but allow ourselves to decide when we have collected enough information from the observations to accept one of the hypotheses with sufficient confidence. A decision rule in this problem consists of two quantities: an $\mathcal{F}_t^Y$-stopping time τ (the decision time) and a $\{0, 1\}$-valued $\mathcal{F}_\tau^Y$-measurable random variable H (the accepted hypothesis, i.e., $H = 1$ means we think that $X = 1$, etc.) We are now faced with the competing goals of minimizing the following quantities: the probability that $H = 1$ when in fact $X = 0$; the probability that $H = 0$ when in fact $X = 1$; and the observation time τ required to determine our accepted hypothesis H. The question is, of course, how to choose (τ, H) to achieve these goals.

Remark 8.4.1. A simple-minded application might help clarify the idea. We wish to send a binary message through a noisy channel using the following communication scheme. At any point in time, we transmit the current bit "telegraph style", i.e., the receiver observes $y_t = \gamma\, X + \sigma\, \xi_t$, where ξ_t is white noise and X is the value of the bit (zero or one). One way of transmitting a message is to allocate a fixed time interval of length Δ for every bit: i.e., we send the first bit during $t \in [0, \Delta[$, the second bit during $[\Delta, 2\Delta[$, etc. The Neyman-Pearson test then provides the optimal way for the receiver to determine the value of each bit in the message, and the probability of error depends purely on Δ. If we thus have an upper bound on the acceptable error probability, we need to choose Δ sufficiently large to attain this bound.

Now suppose, however, that we allow the receiver to signal back to the transmitter when he wishes to start receiving the next bit (e.g., by sending a pulse on a feedback channel). Given a fixed upper bound on the acceptable error probability, we should now be able to decrease significantly the total amount of time necessary to transmit the message. After all, for some realizations of the noise the observations may be relatively unambiguous, while for other realizations it might be very difficult to tell which bit was transmitted. By adapting the transmission time of every bit to the random fluctuations of the noise, we can try to optimize the transmission time of the message while retaining the upper bound on the probability of error. This is a sequential hypothesis testing problem of the type considered in this section.

Bayesian problem

As in the changepoint detection problem, we will begin by solving the "Bayesian" problem and deduce the variational form of the problem at the end of the section. To define the Bayesian problem, we suppose that X is in fact a random variable, independent of B_t, such that $\mathbb{P}(X = 1) = \pi_0$ (where $\pi_0 \in]0, 1[$, otherwise the problem is trivial!). For any decision rule (τ, H), we can then introduce the cost

$$\tilde{J}[\tau, H] = \mathbb{E}(\tau) + a\,\mathbb{P}(X = 1 \text{ and } H = 0) + b\,\mathbb{P}(X = 0 \text{ and } H = 1),$$

where $a > 0$ and $b > 0$ are constants that determine the tradeoff between the two types of error and the length of the observation interval. The goal of the Bayesian problem is to select a decision rule (τ^*, H^*) that minimizes $\tilde{J}[\tau, H]$.

Remark 8.4.2. Nothing is lost by assuming that $\mathbb{E}(\tau) < \infty$, as otherwise the cost is infinite. We will thus always make this assumption throughout this section.

To convert this problem into an optimal stopping problem, our first goal is to eliminate H from the problem. For any fixed stopping rule τ, it is not difficult to find the hypothesis H^*_τ that minimizes $H \mapsto \tilde{J}[\tau, H]$. If we substitute this optimal hypothesis into the cost above, the problem reduces to a minimization of the cost functional $J[\tau] = \tilde{J}[\tau, H^*_\tau]$ over τ only. Let us work out the details.

Lemma 8.4.3. *Denote by π_t the stochastic process with continuous sample paths such that $\pi_t = \mathbb{P}(X = 1|\mathcal{F}^Y_t)$ for every t. Then for any fixed $\mathcal{F}^Y_t$-stopping time τ with $\mathbb{E}(\tau) < \infty$, the cost $\tilde{J}[\tau, H]$ is minimized by accepting the hypothesis*

$$H^*_\tau = \begin{cases} 1 & \text{if } a\pi_\tau \geq b(1 - \pi_\tau), \\ 0 & \text{if } a\pi_\tau < b(1 - \pi_\tau). \end{cases}$$

Moreover, the optimal cost is given by

$$J[\tau] = \tilde{J}[\tau, H^*_\tau] = \mathbb{E}(\tau + a\pi_\tau \wedge b(1 - \pi_\tau)).$$

Proof. As τ is fixed, it suffices to find an $\mathcal{F}^Y_\tau$-measurable H that minimizes $\mathbb{E}(a\,I_{X=1}I_{H=0} + b\,I_{X=0}I_{H=1})$. But using the tower property of the conditional expectation and the optional projection, we find that we can equivalently minimize $\mathbb{E}(a\,\pi_\tau\,(1 - I_{H=1}) + b\,(1 - \pi_\tau)\,I_{H=1})$. But clearly this expression is minimized by H^*_τ, as $a\,\pi_\tau\,(1 - I_{H^*_\tau=1}) + b\,(1 - \pi_\tau)\,I_{H^*_\tau=1} \leq a\,\pi_\tau\,(1 - I_{H=1}) + b\,(1 - \pi_\tau)\,I_{H=1}$ a.s. for any other H. The result now follows directly. □

The filter π_t can be obtained in various ways; we have computed it explicitly in example 7.1.9, and we can easily apply Itô's rule to this expression to obtain a stochastic differential equation. Alternatively, the current case is simply the Shiryaev-Wonham equation with $p_0 + p_\infty = 1$, so we immediately obtain the equation

$$d\pi_t = \frac{\gamma}{\sigma}\,\pi_t(1 - \pi_t)\,d\bar{B}_t, \qquad d\bar{B}_t = \sigma^{-1}(dY_t - \gamma\pi_t\,dt).$$

By the previous lemma, we are seeking a stopping time τ that minimizes

$$J[\tau] = \mathbb{E}\left[\int_0^\tau dt + a\pi_\tau \wedge b(1 - \pi_\tau)\right].$$

This is precisely an optimal stopping problem as formulated in proposition 8.1.3. To solve the problem, we consider as usual the corresponding variational inequality:

$$\min\left\{\frac{\gamma^2 x^2(1-x)^2}{2\sigma^2}\frac{\partial^2 V(x)}{\partial x^2}+1,\ ax\wedge b(1-x)-V(x)\right\}=0.$$

Recall that once we have obtained a suitable solution $V(x)$, the interval $[0,1]$ is divided into the continuation region $D=\{x : V(x) < ax\wedge b(1-x)\}$, on which we must have $\mathscr{L}V(x)+1=0$, and the stopping region $D^c=\{x : V(x)=ax\wedge b(1-x)\}$, on which $\mathscr{L}V(x)+1\geq 0$. Moreover, the function $V(x)$ should be C^1 across the boundary ∂D, if we are going to be able to apply proposition 8.1.3.

We begin by seeking solutions to the equation $\mathscr{L}V(x)+1=0$. This is simply a matter of integrating twice, and we find at once the general solution on $]0,1[$:

$$V_{c,d}(x)=\frac{2\sigma^2}{\gamma^2}(1-2x)\log\left(\frac{x}{1-x}\right)+cx+d.$$

Note that regardless of c,d, the function $V_{c,d}(x)$ is strictly concave and satisfies $V(x)\to-\infty$ as $x\searrow 0$ or $x\nearrow 1$. In particular, this implies that $\partial V_{c,d}(x)/\partial x$ takes every value in $\mathbb{R}$ exactly once on $x\in]0,1[$.

The constants c, d and the continuation region D remain to be found. To this end, we will apply the principle of smooth fit. Let us first narrow down the form of the continuation region. Note that as no $V_{c,d}(x)$ can be made continuous at $x=0$ or $x=1$, the continuation region D must exclude at least some neighborhood of these points. On the other hand, the continuation region must include at least $x=b/(a+b)$, as otherwise $V(x)$ could not be C^1 at this point (why?). Hence the boundary of D must consist of points in the interval $]0,b/(a+b)[$ and in the interval $]b/(a+b),1[$. But on the former, the principle of smooth fit requires that $\partial V(x)/\partial x=a$, while on the latter it must be the case that $\partial V(x)/\partial x=-b$. As the derivative of $V(x)$ takes every value in $\mathbb{R}$ only once, the principle of smooth fit forces the continuation region to be of the form $]\pi^0,\pi^1[$ with $\pi^0\in]0,b/(a+b)[$ and $\pi^1\in]b/(a+b),1[$.

We now need to determine c,d,π^0,π^1. Once we have found π^0, we can directly eliminate c and d; after all, the principle of smooth fit requires that $V(\pi^0)=a\pi^0$ and $\partial V(x)/\partial x|_{x=\pi^0}=a$. Thus, given π^0, we must have

$$V(x)=\Psi(x)-\Psi(\pi^0)+(a-\psi(\pi^0))(x-\pi^0)+a\,\pi^0 \qquad (\text{for } x\in]\pi^0,\pi^1[),$$

where we have written

$$\Psi(x)=\frac{2\sigma^2}{\gamma^2}(1-2x)\log\left(\frac{x}{1-x}\right),\qquad \psi(x)=\frac{\partial\Psi(x)}{\partial x}.$$

The trick is now to select π^0 and π^1 in such a way that $V(x)$ is C^1 at π^1. This gives

$$V(\pi^1)=b(1-\pi^1),\qquad \left.\frac{\partial V(x)}{\partial x}\right|_{x=\pi^1}=\psi(\pi^1)-\psi(\pi^0)+a=-b.$$

We now have two equations relating two unknowns π^0 and π^1; if we can show that a solution exists, then the problem is essentially solved (up to minor technicalities). We will show that there is in fact a unique solution; the proof is illustrated in figure 8.3.

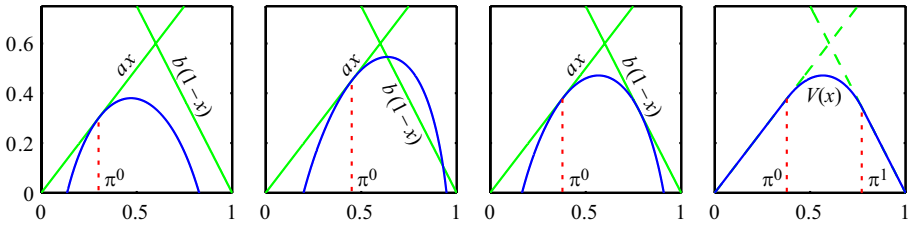

Figure 8.3. Illustration of lemma 8.4.4. If π^0 is chosen too small then $V(x)$ does not touch the line $b(1-x)$ at all (first figure), while if π^0 is too large then the maximum of $V(x) - b(1-x)$ lies above zero (second figure). For the correct choice of π^0, the curve $V(x)$ will be precisely tangent to $b(1-x)$ (third figure). The final construction of the value function (as in theorem 8.4.5) is shown in the last figure. For these plots $\gamma = 1$, $\sigma = .4$, $a = 1$ and $b = 1.5$.

Lemma 8.4.4. *There is a unique π^0, π^1 with $0 < \pi^0 < b/(a+b) < \pi^1 < 1$ such that*

$$\Psi(\pi^1) - \Psi(\pi^0) + (a - \psi(\pi^0))\,(\pi^1 - \pi^0) + a\,\pi^0 = b(1-\pi^1),$$
$$\psi(\pi^0) - \psi(\pi^1) = a + b.$$

Proof. Consider the function $W(x) = V(x) - b\,(1-x)$ (recall that this equation depends on π^0). Note that $W(x)$ is strictly concave for every π^0 and satisfies $W(x) \to -\infty$ as $x \searrow 0$ or $x \nearrow 1$. Hence $W(x)$ has a maximum in the interval $]0,1[$ for every π^0.

For $\pi^0 = b/(a+b)$ the maximum lies above zero: after all, in this case $W(\pi^0) = 0$, while $\partial W(x)/\partial x|_{x=\pi^0} = a$ is positive. As $\pi^0 \to 0$, however, the maximum of $W(x)$ goes below zero. To see this, note that $W(x)$ attains its maximum at the point x^* such that $\partial W(x)/\partial x|_{x=x^*} = \psi(x^*) - \psi(\pi^0) + a + b = 0$, and $\psi(\pi^0) \to \infty$ as $\pi^0 \to 0$; hence $x^* \to 0$ as well. On the other hand, $W(x) \le ax - b\,(1-x)$ everywhere by concavity, so as $x^* \to 0$ we obtain at least $W(x^*) \le -b/2$. Now note that $\psi(x^*) - \psi(\pi^0) + a + b = 0$ implies that x^* is strictly decreasing as π^0 decreases. Hence there must thus be a unique $0 < \pi^0 < b/(a+b)$ such that x^* is precisely zero. But then for that π^0, $V(x^*) = b\,(1-x^*)$ and $\partial V(x)/\partial x|_{x=x^*} = -b$, so we have found the desired π^0 and $\pi^1 = x^*$. Note that $\pi^1 > b/(a+b)$ necessarily, as $V(x) \le ax$ everywhere (so $V(x^*) = b(1-x^*)$ means $b(1-x^*) < ax^*$). □

The remainder of the argument is now routine.

Theorem 8.4.5 (Sequential hypothesis testing). *Define the concave function*

$$V(x) = \begin{cases} ax & \text{for } 0 \le x < \pi^0, \\ \Psi(x) - \Psi(\pi^0) + (a - \psi(\pi^0))\,(x - \pi^0) + a\,\pi^0 & \text{for } \pi^0 \le x \le \pi^1, \\ b(1-x) & \text{for } \pi^1 < x \le 1, \end{cases}$$

where $0 < \pi^0 < \frac{b}{a+b} < \pi^1 < 1$ are the unique points such that $V(\pi^1) = b(1-\pi^1)$ and $\psi(\pi^0) - \psi(\pi^1) = a + b$. Then $V(x)$ is C^1 on $[0,1]$, C^2 on $[0,1]\backslash\{\pi^0, \pi^1\}$, and

$$\min\left\{ \frac{\gamma^2 x^2 (1-x)^2}{2\sigma^2} \frac{\partial^2 V(x)}{\partial x^2} + 1,\ ax \wedge b(1-x) - V(x) \right\} = 0.$$

In particular, the decision strategy

$$\tau^* = \inf\{t : \pi_t \notin]\pi^0, \pi^1[\}, \qquad H^* = \begin{cases} 1 & \text{if } a\pi_{\tau^*} \geq b(1-\pi_{\tau^*}), \\ 0 & \text{if } a\pi_{\tau^*} < b(1-\pi_{\tau^*}), \end{cases}$$

is optimal in that is minimizes the cost $\tilde{J}[\tau, H]$ over all decision rules.

Proof. The various smoothness properties of $V(x)$ hold by construction, while $V(x)$ is easily seen to be concave as its second derivative is nonpositive. Now clearly $\mathscr{L}V(x) + 1 = 0$ on $]\pi^0, \pi^1[$, while on the remainder of the interval $V(x) = ax \wedge b(1-x)$. Moreover, as $V(x)$ is concave and $V(x)$ is tangent to $ax \wedge b(1-x)$ at π^0 and π^1, we must have $V(x) \leq ax \wedge b(1-x)$ everywhere; on the other hand, it is immediately verified that $\mathscr{L}V(x) + 1 \geq 0$ everywhere. Hence the variational inequality is satisfied. We now invoke proposition 8.1.3 with $K = [0,1]$. Clearly $V(x)$ is sufficiently smooth, and as $V(x)$ and both its derivatives are bounded, it remains by lemma 6.3.4 to show that $\mathbb{E}(\tau^*) < \infty$; but this follows immediately from lemma 6.3.3, and the claim is thus established. We are done. □

Variational problem

We now consider the variational version of the problem. Rather than minimizing a cost functional, which trades off between the error probabilities and the length of the observation interval, we now specify fixed upper bounds on the probability of error. We then seek a decision strategy that minimizes the observation time within the class of strategies with acceptable error probabilities. In many situations this is the most natural formulation, and we will see that also this problem has an explicit solution.

We first need to define the problem precisely. To this end, let us denote by $\Delta_{\alpha,\beta}$ the class of decision rules (τ, H) such that

$$\mathbb{P}(H = 0|X = 1) \leq \alpha, \qquad \mathbb{P}(H = 1|X = 0) \leq \beta.$$

For fixed α, β, our goal is to find a decision rule (τ^*, H^*) that minimizes $\mathbb{E}(\tau)$ amongst all decision rules in $\Delta_{\alpha,\beta}$. We will need the following lemmas.

Lemma 8.4.6. *When $\pi_0 \in]\pi^0, \pi^1[$, the rule (τ^*, H^*) of theorem 8.4.5 satisfies*

$$\mathbb{P}(H^* = 0|X = 1) = \frac{\pi^0}{\pi_0}\frac{\pi^1 - \pi_0}{\pi^1 - \pi^0}, \qquad \mathbb{P}(H^* = 1|X = 0) = \frac{1-\pi^1}{1-\pi_0}\frac{\pi_0 - \pi^0}{\pi^1 - \pi^0},$$

Proof. We will consider $\mathbb{P}(H^* = 0|X = 1)$; the remaining claim follows identically. Note that $\mathbb{P}(H^* = 0|X = 1) = \mathbb{P}(H^* = 0 \text{ and } X = 1)/\mathbb{P}(X = 1) = \mathbb{P}(H^* = 0 \text{ and } X = 1)/\pi_0$. Using the tower property of the conditional expectation and the optional projection, we find that $\mathbb{P}(H^* = 0 \text{ and } X = 1) = \mathbb{E}(I_{H^*=0}\pi_{\tau^*}) = \mathbb{E}(I_{\pi_{\tau^*}=\pi^0}\pi_{\tau^*}) = \pi^0\,\mathbb{P}(\pi_{\tau^*} = \pi^0)$. To evaluate the latter, note that π_{τ^*} is a $\{\pi^0, \pi^1\}$-valued random variable; hence

$$\mathbb{E}(\pi_{\tau^*}) = \pi^0\,\mathbb{P}(\pi_{\tau^*} = \pi^0) + \pi^1(1 - \mathbb{P}(\pi_{\tau^*} = \pi^0)) = \pi_0 \quad \Longrightarrow \quad \mathbb{P}(\pi_{\tau^*} = \pi^0) = \frac{\pi^1 - \pi_0}{\pi^1 - \pi^0},$$

as π_t is a bounded martingale. This establishes the result. □

Lemma 8.4.7. *Given $0 < \alpha + \beta < 1$ and $\pi_0 \in]0,1[$, there are unique constants $a, b > 0$ in the Bayesian cost $\tilde{J}[\tau, H]$ such that (τ^*, H^*) of theorem 8.4.5 satisfies $\mathbb{P}(H^* = 0|X = 1) = \alpha$ and $\mathbb{P}(H^* = 1|X = 0) = \beta$; moreover, for these a, b we find*

$$\pi^0 = \frac{\pi_0 \alpha}{(1-\pi_0)(1-\beta) + \pi_0 \alpha}, \qquad \pi^1 = \frac{\pi_0(1-\alpha)}{(1-\pi_0)\beta + \pi_0(1-\alpha)},$$

where it is easily verified that $0 < \pi^0 < \pi_0 < \pi^1 < 1$.

Proof. The π^0 and π^1 in the lemma are obtained by setting $\mathbb{P}(H^* = 0|X = 1) = \alpha$ and $\mathbb{P}(H^* = 1|X = 0) = \beta$ in the previous lemma, then solving for π^0 and π^1. It is easily verified that $0 < \alpha + \beta < 1$ and $\pi_0 \in]0,1[$ ensures that $0 < \pi^0 < \pi_0 < \pi^1 < 1$. It remains to find a, b that give rise to these π^0, π^1; but substituting π^0, π^1 into lemma 8.4.4, we find a linear system of equations for a, b which clearly has a unique solution. Thus the claim is established. □

We can now proceed to solve the variational problem as in corollary 8.3.6.

Lemma 8.4.8. *For any $\pi_0 \in]0,1[$, the optimal Bayesian decision rule (τ^*, H^*) with a, b as in the previous lemma is optimal for the variational problem.*

Proof. Note that $(\tau^*, H^*) \in \Delta_{\alpha,\beta}$ by construction. It remains to show that for any $(\tau, H) \in \Delta_{\alpha,\beta}$, we have $\mathbb{E}(\tau^*) \leq \mathbb{E}(\tau)$. But as (τ^*, H^*) is optimal for the Bayesian problem,

$$\mathbb{E}(\tau^*) + a\pi_0\alpha + b(1-\pi_0)\beta = \tilde{J}[\tau^*, H^*] \leq \tilde{J}[\tau, H] \leq \mathbb{E}(\tau) + a\pi_0\alpha + b(1-\pi_0)\beta,$$

so it is indeed the case that $\mathbb{E}(\tau^*) \leq \mathbb{E}(\tau)$. This establishes the claim. □

Though this result does, in principle, solve the variational problem, the true structure of the problem is still in disguise. It is illuminating to remove the somewhat strange dependence of the stopping boundaries π^0, π^1 on $\pi_0 = \mathbb{P}(X = 1)$ through a change of variables. To this end, let us define the likelihood ratio

$$\varphi_t \equiv \frac{\pi_t}{1-\pi_t}\,\frac{1-\pi_0}{\pi_0} = \exp\left(\frac{\gamma}{\sigma^2}\,Y_t - \frac{\gamma^2}{2\sigma^2}\,t\right),$$

where the latter equality can be read off from example 7.1.9. As $x/(1-x)$ is strictly increasing, the stopping rule (τ^*, H^*) of lemma 8.4.8 can be equivalently written as

$$\tau^* = \inf\left\{t : \varphi_t \notin \left]\frac{\alpha}{1-\beta}, \frac{1-\alpha}{\beta}\right[\right\}, \qquad H^* = \begin{cases} 1 & \text{if } \varphi_{\tau^*} \geq (1-\alpha)/\beta, \\ 0 & \text{if } \varphi_{\tau^*} \leq \alpha/(1-\beta). \end{cases}$$

Evidently the optimal variational decision rule (τ^*, H^*) can be computed without any knowledge of π_0; after all, both the functional φ_t and the stopping boundaries no longer depend on π_0. Hence we find that unlike in the Bayesian problem, *no probabilistic assumption needs to be made on the law of X in the variational problem.* Indeed, we can consider the value of X simply as being "unknown", rather than "random". This is very much in the spirit of the Neyman-Pearson test, and the two methods are in fact more closely related than our approach indicates (see [Shi73, section IV.2]).

8.5 Impulse control

In many ways optimal stopping problems *are* optimal control problems—they involve the choice of a strategy which, if followed, achieves a particular goal; this is in essence what control is all about! However, techniques very similar to the ones used in optimal stopping theory can also be useful in conjunction with more traditional types of control. For example, one could investigate a combination of the indefinite cost control problem of chapter 6 and an optimal stopping problem, where the goal is to optimize simultaneously over the time at which the control is terminated and the continuous control strategy followed up to that time. A description of such combined problems can be found, for example, in [ØS05, chapter 4].

In this final section we will discuss a different combination of optimal control and optimal stopping techniques. We are interested in the situation where there is no terminal time—we will control the system of interest on the infinite time horizon. However, unlike in chapter 6, where we applied a control continuously in time, we will only apply a control action at a sequence of stopping times, i.e., we allow ourselves to give the system impulses at suitably chosen times; this is called an *impulse control problem*. Such problems are important in a variety of applications, including resource management (when should we cut down and replant a forest to maximize the yield?), inventory management (when should we restock our warehouse?), production planning (when to start and stop production?), and economic applications.[2]

Before we can solve the control problem, we need to make precise what we mean by an impulse control strategy. For times prior to the first intervention time $t < \tau_1$, the system evolves according to the stochastic differential equation

$$X_t^u = X_0 + \int_0^t b(X_s^u)\,ds + \int_0^t \sigma(X_s^u)\,dW_s,$$

where X_0 is $\mathcal{F}_0$-measurable and $b : \mathbb{R}^n \to \mathbb{R}^n$, $\sigma : \mathbb{R}^n \to \mathbb{R}^{n\times m}$ satisfy appropriate conditions that ensure existence and uniqueness of the solution. At the stopping time τ_1, we impulsively change the system state from $X^u_{\tau_1-}$ to $X^u_{\tau_1} = \Gamma(X^u_{\tau_1-}, \zeta_1)$, where $\Gamma : \mathbb{R}^n \times \mathbb{U} \to \mathbb{R}^n$ is a given control action function and $\mathbb{U}$ is the control set. The control ζ_1 is assumed to be $\mathcal{F}_{\tau_1}$-measurable, i.e., the control strategy is adapted.

Remark 8.5.1. As the state of the system jumps at the intervention time, we need to have notation that distinguishes between the state just prior and just after the intervention. In the following, we will denote by $X_{\tau-}$ the state just prior to the intervention time τ, and by X_τ the state just after the intervention time τ. We will thus always have $X_\tau = \Gamma(X_{\tau-}, \zeta)$, where ζ is the control applied at time τ.

[2] An interesting economic application is the following. Due to various economic factors, the exchange rate between two currencies (say, the dollar and some foreign currency) fluctuates randomly in time. It is not a good idea, however, to have the exchange rate be too far away from unity. As such, the central bank tries to exert its influence on the exchange rates to keep them in a certain "safe" target zone. One way in which the central bank can influence the exchange rate is by buying or selling large quantities of foreign currency. The question then becomes, at which points in time should the central bank decide to make a large transaction in foreign currency, and for what amount, in order to keep the exchange rate in the target zone. This is an impulse control problem. See [Kor99] for a review of impulse control applications in finance.

We left off just after the first intervention time τ_1. For times after τ_1 but before the second intervention time $\tau_2 > \tau_1$, we solve the stochastic differential equation

$$X_t^u = X_{\tau_1}^u + \int_{\tau_1}^t b(X_s^u)\,ds + \int_{\tau_1}^t \sigma(X_s^u)\,dW_s.$$

We will assume that such an equation has a unique solution starting from every finite stopping time, so that we can solve for X_t between every pair $\tau_i < \tau_{i+1}$.

Remark 8.5.2. This is indeed the case when b, σ satisfy the usual Lipschitz conditions; this follows from the strong Markov property, but let us not dwell on this point.

At time τ_2 we apply another control action ζ_2, etc. We now have the following.

Definition 8.5.3. An *impulse control strategy* u consists of

1. a sequence of stopping times $\{\tau_j\}_{j=1,2,\ldots}$ such that $\tau_j < \infty$ a.s. and $\tau_j < \tau_{j+1}$;
2. a sequence $\{\zeta_j\}_{j=1,2,\ldots}$ such that $\zeta_j \in \mathbb{U}$ and ζ_j is $\mathcal{F}_{\tau_j}$-measurable.

The strategy u is called *admissible* if the intervention times τ_j do not accumulate, and X_t^u has a unique solution on the infinite time interval $[0, \infty[$.

Let us investigate the discounted version of the impulse control problem (a time-average cost can also be investigated, you can try to work this case out yourself or consult [JZ06]). We introduce the following discounted cost functional:

$$J[u] = \mathbb{E}\left[\int_0^\infty e^{-\lambda s}\, w(X_s)\,ds + \sum_{j=1}^\infty e^{-\lambda \tau_j}\, v(X_{\tau_j -}, \zeta_j)\right].$$

Here $w : \mathbb{R}^n \to \mathbb{R}$ is the running cost, $v : \mathbb{R}^n \times \mathbb{U} \to \mathbb{R}$ is the intervention cost, and $\lambda > 0$ is the discounting factor. We seek an admissible impulse control strategy u^* that minimizes the cost $J[u]$. To this end we will prove a verification theorem.

Proposition 8.5.4. *Assume that w and v are either both bounded from below, or both bounded from above. Let $K \subset \mathbb{R}^n$ be a set such that $X_t \in K$ for all t, and suppose there is a $V : K \to \mathbb{R}$, which is sufficiently smooth to apply Itô's rule, such that*

$$\min\{\mathscr{L}V(x) - \lambda\, V(x) + w(x),\ \mathscr{K}V(x) - V(x)\} = 0,$$

where the intervention operator $\mathscr{K}$ is defined as

$$\mathscr{K}V(x) = \min_{\alpha \in \mathbb{U}}\{V(\Gamma(x,\alpha)) + v(x,\alpha)\}.$$

Assume that $|\mathbb{E}(V(X_0))| < \infty$, and denote by $\mathfrak{K}$ the class of admissible strategies u such that $\mathbb{E}(e^{-\lambda\tau_j} V(X^u_{\tau_j -})) \xrightarrow[j\to\infty]{} 0$ and such that

$$\mathbb{E}\left[\sum_{i=1}^n \sum_{k=1}^m \int_{\tau_{j-1}}^{\tau_j} e^{-\lambda s}\, \frac{\partial V}{\partial x^i}(X_s^u)\, \sigma^{ik}(X_s^u)\, dW_s^k\right] = 0 \quad \textit{for all } j.$$

Define the continuation set $D = \{x \in K : \mathscr{K}V(x) > V(x)\}$ *and strategy* u^* *with*

$$\tau_j^* = \inf\{t > \tau_{j-1}^* : X_t^{u^*} \notin D\}, \quad \zeta_j^* \in \operatorname{argmin}_{\alpha\in\mathbb{U}}\{V(\Gamma(X_{\tau_j-}^{u^*}, \alpha)) + v(X_{\tau_j-}^{u^*}, \alpha)\}.$$

If u^* *defines an admissible impulse control strategy in* $\mathfrak{K}$*, then* $J[u^*] \le J[u]$ *for any* $u \in \mathfrak{K}$*, and the optimal cost can be written as* $\mathbb{E}(V(X_0)) = J[u^*]$.

Proof. We may assume without loss of generality that w and v are both nonnegative or both nonpositive; otherwise this can always be accomplished by shifting the cost by a constant. We now begin by applying Itô's rule to $e^{-\lambda t}V(X_t^u)$. Familiar manipulations give

$$\mathbb{E}(e^{-\lambda\tau_{n-1}}V(X_{\tau_{n-1}}^u)) - \mathbb{E}(e^{-\lambda\tau_n}V(X_{\tau_n-}^u)) = \mathbb{E}\left[\int_{\tau_{n-1}}^{\tau_n} e^{-\lambda s}\{\lambda V(X_s^u) - \mathscr{L}V(X_s^u)\}\,ds\right] \le \mathbb{E}\left[\int_{\tau_{n-1}}^{\tau_n} e^{-\lambda s}\,w(X_s^u)\,ds\right].$$

Summing n from 1 to j (set $\tau_0 = 0$), we obtain

$$\mathbb{E}(V(X_0)) - \mathbb{E}(e^{-\lambda\tau_j}V(X_{\tau_j-}^u)) \le \mathbb{E}\left[\int_0^{\tau_j} e^{-\lambda s}\,w(X_s^u)\,ds + \sum_{i=1}^{j-1} e^{-\lambda\tau_i}(V(X_{\tau_i-}) - V(X_{\tau_i}))\right].$$

But $V(X_{\tau_i}) = V(\Gamma(X_{\tau_i-}, \zeta_i)) \ge \mathscr{K}V(X_{\tau_i-}) - v(X_{\tau_i-}, \zeta_i)$ by the definition of the intervention operator, so $V(X_{\tau_i-}) - V(X_{\tau_i}) \le V(X_{\tau_i-}) - \mathscr{K}V(X_{\tau_i-}) + v(X_{\tau_i-}, \zeta_i) \le v(X_{\tau_i-}, \zeta_i)$ using the fact that $\mathscr{K}V(x) - V(x) \ge 0$. Hence

$$\mathbb{E}(V(X_0)) - \mathbb{E}(e^{-\lambda\tau_j}V(X_{\tau_j-}^u)) \le \mathbb{E}\left[\int_0^{\tau_j} e^{-\lambda s}\,w(X_s^u)\,ds + \sum_{i=1}^{j-1} e^{-\lambda\tau_i}\,v(X_{\tau_i-}, \zeta_i)\right].$$

Now let $t, j \to \infty$, using monotone convergence on the right and the assumption on $u \in \mathfrak{K}$ on the left (recall that as u is admissible, the intervention times cannot accumulate so $\tau_j \nearrow \infty$). This gives $\mathbb{E}(V(X_0)) \le J[u]$. Repeating the same arguments with u^* instead of u gives $\mathbb{E}(V(X_0)) = J[u^*]$, so the claim is established. □

Remark 8.5.5. The equation for the value function $V(x)$ is almost a variational inequality, but not quite; in a variational inequality, the stopping cost $z(x)$ was independent of $V(x)$, while in the current problem the intervention cost $\mathscr{K}V(x)$ very much depends on the value function (in a nontrivial manner!). The equation for $V(x)$ in proposition 8.5.4 is known as a *quasivariational inequality*.

Let us treat an interesting example, taken from [Wil98].

Example 8.5.6 (Optimal forest harvesting). We own a forest which is harvested for lumber. When the forest is planted, it starts off with a (nonrandom) total biomass $x_0 > 0$; as the forest grows, the biomass of the forest grows according the equation

$$dX_t = \mu X_t\,dt + \sigma X_t\,dW_t, \quad X_0 = x_0 \qquad (\mu > 0).$$

At some time τ_1, we can decide to cut the forest and sell the wood; we then replant the forest so that it starts off again with biomass x_0. The forest can then grow freely

until we decide to cut and replant again at time τ_2, etc. Every time τ we cut the forest, we obtain $X_{\tau-}$ dollars from selling the wood, but we pay a fee proportional to the total biomass $\alpha X_{\tau-}$ ($0 \leq \alpha < 1$) for cutting the forest, and a fixed fee $Q > 0$ for replanting the forest to its initial biomass x_0. When inflation, with rate[3] $\lambda > \mu$, is taken into account, the expected future profit from this operation is given by

$$\mathbb{E}\left[\sum_{j=1}^{\infty} e^{-\lambda\tau_j}\left((1-\alpha)\,X_{\tau_j-} - Q\right)\right].$$

Our goal is to choose a harvesting strategy $\tau_1, \tau_2, \ldots$ which maximizes our expected profit, i.e., we wish to choose an impulse control strategy u^* which minimizes

$$J[u] = \mathbb{E}\left[\sum_{j=1}^{\infty} e^{-\lambda\tau_j}\left(Q - (1-\alpha)\,X_{\tau_j-}\right)\right].$$

For this impulse control problem, $\mathbb{U}$ consists of only one point (so we can essentially ignore it) and the control action is $\Gamma(x,\alpha) = x_0$ for any x. Note, moreover, that $X_t > 0$ always, so we can apply proposition 8.5.4 with $K =]0,\infty[$.

To solve the impulse control problem, we consider the quasivariational inequality

$$\min\left\{\frac{\sigma^2x^2}{2}\frac{\partial^2V(x)}{\partial x^2} + \mu x\,\frac{\partial V(x)}{\partial x} - \lambda\,V(x),\; Q - \beta x + V(x_0) - V(x)\right\} = 0,$$

where $\beta = 1-\alpha$. The first thing to note is that in order to obtain a meaningful impulse control strategy, the initial biomass x_0 must be in the continuation set D; if this is not the case, then replanting the forest is so cheap that you might as well immediately cut down what has just been planted, without waiting for it to grow. To avoid this possibility, note that $x_0 \in D$ requires $x_0 < Q/\beta$. We will assume this from now on.

Now consider $\mathscr{L}V(x) - \lambda V(x) = 0$, which must hold on the continuation region D. The general solution to this equation is given by $V(x) = c_+x^{\gamma_+} + c_-x^{\gamma_-}$, where

$$\gamma_\pm = \frac{\sigma^2 - 2\mu \pm \sqrt{(\sigma^2-2\mu)^2 + 8\sigma^2\lambda}}{2\sigma^2}.$$

Note that $\gamma_+ > 1$ (due to $\lambda > \mu$), while $\gamma_- < 0$.

We could proceed to analyze every possible case, but let us make a few educated guesses at this point. There is nothing lost by doing this: if we can find one solution that satisfies the conditions of the verification theorem, then we are done; otherwise we can always go back to the drawing board! We thus guess away. First, it seems unlikely that it will be advantageous to cut the forest when there is very little biomass; this will only cause us to pay the replanting fee Q, without any of the benefit of selling the harvested wood. Hence we conjecture that the continuation region has the form $D =]0, y[$ for some $y > x_0$. In particular, this means the $V(x) = c_+x^{\gamma_+} + c_-x^{\gamma_-}$ for

[3] If the inflation rate were lower than the mean growth rate of the forest, then it never pays to cut down the forest—if we are patient, we can always make more money by waiting longer before cutting the forest.

$x \in]0, y[$. But as the second term has a pole at zero, we must choose $c_- = 0$; after all, $c_- > 0$ is impossible as the cost is bounded from above, while $c_- < 0$ would imply that we make more and more profit the less biomass there is in the forest; clearly this cannot be true. Collecting these ideas, we find that $V(x)$ should be of the form

$$V(x) = \begin{cases} c\,x^{\gamma_+} & \text{for } x < y, \\ Q - \beta x + c\,x_0^{\gamma_+} & \text{for } x \geq y. \end{cases}$$

The constants c and y remain to be determined. To this end, we apply the principle of smooth fit. As $V(x)$ should be C^1 at y, we require

$$\gamma_+ c\,y^{\gamma_+ - 1} = -\beta, \qquad c\,y^{\gamma_+} = Q - \beta y + c\,x_0^{\gamma_+}.$$

Hence we obtain the candidate value function

$$V(x) = \begin{cases} \psi(x) & \text{for } x < y, \\ Q - \beta x + \psi(x_0) & \text{for } x \geq y, \end{cases} \qquad \psi(x) = -\frac{\beta y}{\gamma_+}\left(\frac{x}{y}\right)^{\gamma_+},$$

where $y > x_0$ solves the equation

$$y = \frac{\gamma_+ Q - \beta y\,(x_0/y)^{\gamma_+}}{\beta(\gamma_+ - 1)}.$$

To complete the argument, it remains to show (i) that there does exist a solution $y > x_0$; and (ii) that the conditions of proposition 8.5.4 are satisfied for $V(x)$.

Let us first deal with question (i). Let $f(z) = \beta(\gamma_+ - 1)z + \beta z\,(x_0/z)^{\gamma_+} - \gamma_+ Q$; then y satisfies $f(y) = 0$. It is easily verified that $f(z)$ is strictly convex and attains its minimum at $z = x_0$; furthermore, $f(x_0) < 0$ as we have assumed that $x_0 < Q/\beta$. Hence $f(z)$ has exactly two roots, one of which is larger than x_0. Hence we find that there exists a unique $y > x_0$ that satisfies the desired relation.

We now verify (ii). Note that $V(x)$ is, by construction, C^1 on $]0, \infty[$ and C^2 on $]0, \infty[\setminus\{y\}$. Hence $V(x)$ is sufficiently smooth. The running cost w is zero in our case, while the intervention cost v is bounded from above. By construction, $\mathscr{L}V(x) - \lambda V(x) = 0$ on $D =]0, y[$, while $V(x) = \mathscr{K}V(x)$ on D^c. It is easily verified by explicit computation that $\mathscr{L}V(x) - \lambda V(x) \geq 0$ on D^c and that $V(x) < \mathscr{K}V(x)$ on D. Hence the quasivariational inequality is satisfied. It thus remains to show that the candidate optimal strategy u^* is admissible, and in particular that it is in $\mathfrak{K}$.

To show that this is the case, we proceed as follows. First, we claim that $\tau_j < \infty$ a.s. for every j. To see this, it suffices to note that for the uncontrolled process $\mathbb{E}(X_t) = x_0\,e^{\mu t} \to \infty$, so there exists a subsequence t_n such that $X_{t_n} \to \infty$ a.s., so that in particular X_t must eventually exit D a.s. Furthermore, due to the continuity of the sample paths of X_t, we can immediately see that $\tau_{j-1} < \tau_j$. Now note that our process X_t^u restarts at the same, non-random point at every intervention time τ_j. In particular, this means that $\tau_j - \tau_{j-1}$ are independent of each other for every j, and as $\tau_j - \tau_{j-1} > 0$, there must exist some $\varepsilon > 0$ such that $\mathbb{P}(\tau_j - \tau_{j-1} > \varepsilon) > 0$. These two facts together imply that $\mathbb{P}(\tau_j - \tau_{j-1} > \varepsilon \text{ i.o.}) = 1$ (see, for example, the argument in the example at the end of section 4.1). But then we conclude that the stopping times τ_j cannot accumulate, and in particular $\tau_j \nearrow \infty$. Thus u^* is admissible.

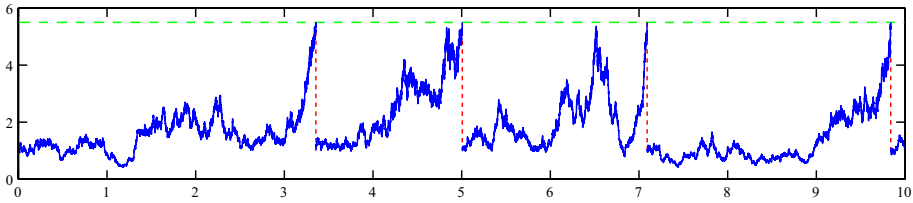

Figure 8.4. Simulation of the optimal impulse control strategy of example 8.5.6. One sample path of the controlled process X_t^u is shown in blue; the intervention threshold y is shown in green. The parameters for this simulation were $\mu = \sigma = 1$, $\lambda = Q = 2$, $\alpha = .1$, and $x_0 = 1$.

To show that u^* is also in $\mathfrak{K}$ is now not difficult. Indeed, as D has compact closure, X_t^u is a bounded process. Hence $\mathbb{E}(e^{-\lambda\tau_j} V(X^u_{\tau_j -})) \to 0$, while

$$\mathbb{E}\left[\sum_{i=1}^{n}\sum_{k=1}^{m}\int_{\tau_{j-1}}^{\tau_j} e^{-\lambda s}\,\frac{\partial V}{\partial x^i}(X_s^u)\,\sigma^{ik}(X_s^u)\,dW_s^k\right] = 0$$

follows from the fact that the integrand is square-integrable on the infinite time horizon (being the product of a decaying exponential and a bounded process). Thus all the requirements of proposition 8.5.4 are satisfied, and we are convinced that we have indeed found an optimal impulse control strategy, as we set out to do (see figure 8.4).

More elaborate optimal harvesting models can be found in [Wil98] and in [Alv04].

8.6 Further reading

The recent book by Peskir and Shiryaev [PS06] is an excellent resource on optimal stopping theory, and covers in depth the fundamental theory, methods of solution, and a wide range of applications; if you wish to learn more about optimal stopping, this book is very highly recommended. Shiryaev's older monograph [Shi73] is also still a classic on the topic. Both these references develop in detail the connections between optimal stopping problems and so-called free boundary problems, which is what is obtained when the variational inequality is combined with the principle of smooth fit. The name should be obvious: these are PDEs whose boundary conditions live on a "free" boundary, which is itself a part of the solution. For much more on this topic, see Bensoussan and Lions [BL82, Ben82, BL84] and Friedman [Fri75]. Øksendal [Øks03] and Øksendal and Sulem [ØS05] contain some useful verification theorems. The fundamental theory of optimal stopping is still being developed after all these years; see, e.g., Dayanik and Karatzas [DK03] for a recent contribution.

Numerical methods for optimal stopping problems, using Markov chain approximations, are detailed in Kushner [Kus77] and Kushner and Dupuis [KD01]. A discussion of optimal stopping for discrete time, discrete state space Markov chains can be found in [Kus71]; a nice introduction to this topic is also given in [Bil86].

An in-depth study of the optional projection and friends (who we did not introduce) can be found in Dellacherie and Meyer [DM82]. For more on the separation principle in the optimal stopping setting see, e.g., Szpirglas and Mazziotto [SM79].

Our treatment of both the changepoint detection problem and the hypothesis testing problem come straight from Shiryaev [Shi73], see also [PS06]. For more on the expected miss criterion see Karatzas [Kar03], while the stock selling problem is from Rishel and Helmes [RH06], where the risk-neutral version can also be found. Many interesting applications of changepoint detection can be found in Basseville and Nikiforov [BN93]. An extension of the hypothesis testing problem to time-varying signals can be found in Liptser and Shiryaev [LS01b, section 17.6].

Our discussion of the impulse control problem is inspired by Øksendal and Sulem [ØS05] and by Brekke and Øksendal [BØ94]. An extension of example 8.1.7 to the impulse control setting can be found in the latter. The time-average cost criterion is discussed, e.g., in Jack and Zervos [JZ06]. Finally, the classic tome on quasivariational inequalities, and a rich source of examples of impulse control applications in management problems, is Bensoussan and Lions [BL84]. Markov chain approximations for the solution of impulse control problems can be found in Kushner [Kus77].

APPENDIX A

Problem sets

A.1 Problem set 1

Q. 1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space on which is defined a sequence of i.i.d. Gaussian random variables $\xi_1, \xi_2, \ldots$ with zero mean and unit variance. Consider the following recursion:

$$x_n = e^{a+b\xi_n} x_{n-1}, \qquad x_0 = 1,$$

where a and b are real-valued constants. This is a crude model for some nonnegative quantity that grows or shrinks randomly in every time step; for example, we could model the price of a stock this way, albeit in discrete time.

1. Under which conditions on a and b do we have $x_n \to 0$ in $\mathcal{L}^p$?
2. Show that if $x_n \to 0$ in $\mathcal{L}^p$ for some $p > 0$, then $x_n \to 0$ a.s.

   Hint: prove $x_n \to 0$ in $\mathcal{L}^p \implies x_n \to 0$ in probability $\implies x_n \to 0$ a.s.
3. Show that if there is no $p > 0$ s.t. $x_n \to 0$ in $\mathcal{L}^p$, then $x_n \not\to 0$ in any sense.
4. If we interpret x_n as the price of stock, then x_n is the amount of dollars our stock is worth by time n if we invest one dollar in the stock at time 0. If $x_n \to 0$ a.s., this means we eventually lose our investment with unit probability. However, it is possible for a and b to be such that $x_n \to 0$ a.s., but nonetheless our *expected winnings* $\mathbb{E}(x_n) \to \infty$! Find such a, b. Would you consider investing in such a stock? *[Any answer is acceptable, as long as it is well motivated.]*

Q. 2. We work on the probability space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P})$, where the probability measure $\mathbb{P}$ is such that the canonical random variable $X : \omega \mapsto \omega$ is a Gaussian random variable with zero mean and unit variance. In addition to $\mathbb{P}$, we consider a probability measure

$\mathbb{Q}$ under which $X - a$ is a Gaussian random variable with zero mean and unit variance, where $a \in \mathbb{R}$ is some fixed (non-random) constant.

1. Is it true that $\mathbb{Q} \ll \mathbb{P}$, and if so, what is the Radon-Nikodym derivative $d\mathbb{Q}/d\mathbb{P}$? Similarly, is it true that $\mathbb{P} \ll \mathbb{Q}$, and if so, what is $d\mathbb{P}/d\mathbb{Q}$?

We are running a nuclear reactor. That being a potentially dangerous business, we would like to detect the presence of a radiation leak, in which case we should shut down the reactor. Unfortunately, we only have a noisy detector: the detector generates some random value ξ when everything is ok, while in the presence of a radiation leak the noise has a constant offset $a + \xi$. Based on the value returned by the detector, we need to make a decision as to whether to shut down the reactor.

In our setting, the value returned by the detector is modelled by the random variable X. If everything is running ok, then the outcomes of X are distributed according to the measure $\mathbb{P}$. This is called the *null hypothesis* H_0. If there is a radiation leak, however, then X is distributed according to $\mathbb{Q}$. This is the *alternative hypothesis* H_1. Based on the value X returned by the detector, we decide to shut down the reactor if $f(X) = 1$, with some $f : \mathbb{R} \to \{0, 1\}$. Our goal is to find a suitable function f.

How do we choose the decision function f? What we absolutely cannot tolerate is that a radiation leak occurs, but we do not decide to shut down the reactor—disaster would ensue! For this reason, we fix a tolerance threshold: under the measure corresponding to H_1, the probability that $f(X) = 0$ must be at most some fixed value α (say, 10^{-12}). That is, we insist that any acceptable f must be such that $\mathbb{Q}(f(X) = 0) \leq \alpha$. Given this constraint, we now try to find an acceptable f that minimizes $\mathbb{P}(f(X) = 1)$, the probability of *false alarm* (i.e., there is no radiation leak, but we think there is).

Claim: an f^* that minimizes $\mathbb{P}(f(X) = 1)$ subject to $\mathbb{Q}(f(X) = 0) \leq \alpha$ is

$$f^*(x) = \begin{cases} 1 & \text{if } \frac{d\mathbb{Q}}{d\mathbb{P}}(x) > \beta, \\ 0 & \text{otherwise,} \end{cases}$$

where $\beta > 0$ is chosen such that $\mathbb{Q}(f^*(X) = 0) = \alpha$. This is called the *Neyman-Pearson test*, and is a very fundamental result in statistics (if you already know it, all the better!). You are going to prove this result.

2. Let $f : \mathbb{R} \to \{0, 1\}$ be an arbitrary measurable function s.t. $\mathbb{Q}(f(X) = 0) \leq \alpha$. Using $\mathbb{Q}(f(X) = 0) \leq \alpha$ and $\mathbb{Q}(f^*(X) = 0) = \alpha$, show that

$$\mathbb{Q}(f^*(X) = 1 \text{ and } f(X) = 0) \leq \mathbb{Q}(f^*(X) = 0 \text{ and } f(X) = 1).$$

3. Using the definition of f^*, show that the previous inequality implies

$$\mathbb{P}(f^*(X) = 1 \text{ and } f(X) = 0) \leq \mathbb{P}(f^*(X) = 0 \text{ and } f(X) = 1).$$

   Finally, complete the proof of optimality of the Neyman-Pearson test by adding a suitable quantity to both sides of this inequality.

A better detector would give a sequence $X_1, \ldots, X_N$ of measurements. Under the measure $\mathbb{P}$ (everything ok), the random variables $X_1, \ldots, X_N$ are independent Gaussian random variables with zero mean and unit variance; under the measure $\mathbb{Q}$ (radiation leak), the random variables $X_1 - a_1, \ldots, X_N - a_N$ are independent Gaussian random variables with zero mean and unit variance, where $a_1, \ldots, a_N$ is a fixed (non-random) alarm signal (for example, a siren $a_n = \sin(n\pi/2)$.)

4. Construct $X_1, \ldots, X_N$, $\mathbb{P}$ and $\mathbb{Q}$ on a suitable product space. What is $d\mathbb{Q}/d\mathbb{P}$? How does the Neyman-Pearson test work in this context?

5. **Bonus question:** Now suppose that we have an entire sequence $X_1, X_2, \ldots$, which are i.i.d. Gaussian random variables with mean zero and unit variance under $\mathbb{P}$, and such that $X_1 - a_1, X_2 - a_2, \ldots$ are i.i.d. Gaussian random variables with mean zero and unit variance under $\mathbb{Q}$. Give a necessary and sufficient condition on the non-random sequence $a_1, a_2, \ldots$ so that $\mathbb{Q} \ll \mathbb{P}$. In the case that $\mathbb{Q} \ll \mathbb{P}$, give the corresponding Radon-Nikodym derivative. If $\mathbb{Q} \not\ll \mathbb{P}$, find an event A so that $\mathbb{P}(A) = 0$ but $\mathbb{Q}(A) \neq 0$. In theory, how would you solve the hypothesis testing problem when $\mathbb{Q} \ll \mathbb{P}$? How about when $\mathbb{Q} \not\ll \mathbb{P}$?

A.2 Problem set 2

Q. 3. Let W_t be a Wiener process.

1. Prove that $\tilde{W}_t = cW_{t/c^2}$ is also a Wiener process for any $c > 0$. Hence the sample paths of the Wiener process are *self-similar* (or *fractal*).

2. Define the stopping time $\tau = \inf\{t > 0 : W_t = x\}$ for some $x > 0$. Calculate the moment generating function $\mathbb{E}(e^{-\lambda\tau})$, $\lambda > 0$ by proceding as follows:

   a) Prove that $X_t = e^{(2\lambda)^{1/2}W_t - \lambda t}$ is a martingale. Show that $X_t \to 0$ a.s. as $t \to \infty$ (first argue that X_t converges a.s.; it then suffices to show that $X_n \to 0$ a.s. ($n \in \mathbb{N}$), for which you may invoke Q.1 in homework 1.)

   b) It follows that $Y_t = X_{t\wedge\tau}$ is also a martingale. Argue that Y_t is bounded, i.e., $Y_t < K$ for some $K > 0$ and all t, and that $Y_t \to X_\tau$ a.s. as $t \to \infty$.

   c) Show that it follows that $\mathbb{E}(X_\tau) = 1$ (this is almost the optional stopping theorem, except that we have not required that $\tau < \infty$!) The rest is easy.

   What is the mean and variance of τ? (You don't have to give a rigorous argument.) In particular, does W_t always hit the level x in finite time?

Q. 4 (Lyapunov functions). In deterministic nonlinear systems and control theory, the notions of (Lyapunov) *stability*, *asymptotic stability*, and *global stability* play an important role. To prove that a system is stable, one generally looks for a suitable *Lyapunov function*, as you might have learned in a nonlinear systems class. Our goal is to find suitable stochastic counterparts of these ideas, albeit in discrete time.

We work on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which is defined a sequence of i.i.d. random variables $\xi_1, \xi_2, \ldots$ We consider a dynamical system defined by the recursion

$$x_n = F(x_{n-1}, \xi_n) \quad (n = 1, 2, \ldots), \qquad x_0 \text{ is non-random,}$$

where $F : S \times \mathbb{R} \to S$ is some *continuous* function and S is some compact subset of $\mathbb{R}^d$ (compactness is not essential, but we go with it for simplicity). Let us assume that $F(x^*, \xi) = x^*$ for some $x^* \in S$ and all $\xi \in \mathbb{R}$.

The following notions of stability are natural counterparts of the deterministic notions (see your favorite nonlinear systems textbook). The equilibrium x^* is

- **stable** if for any $\varepsilon > 0$ and $\alpha \in]0, 1[$, there exists a $\delta < \varepsilon$ such that we have $\mathbb{P}(\sup_{n \geq 0} \|x_n - x^*\| < \varepsilon) > \alpha$ whenever $\|x_0 - x^*\| < \delta$ *("if we start close to x^*, then with high probability we will remain close to x^* forever");*
- **asymptotically stable** if it is stable and for every $\alpha \in]0, 1[$, there exists a κ such that $\mathbb{P}(x_n \to x^*) > \alpha$ whenever $\|x_0 - x^*\| < \kappa$ *("if we start sufficiently close to x^*, then we will converge to x^* with high probability");*
- **globally stable** if it is stable and $x_n \to x^*$ a.s. for any x_0.

1. Prove the following theorem:

   Theorem A.2.1. *Suppose that there is a continuous function $V : S \to [0, \infty[$, with $V(x^*) = 0$ and $V(x) > 0$ for $x \neq x^*$, such that*

   $$\mathbb{E}(V(F(x, \xi_n))) - V(x) = k(x) \leq 0 \quad \textit{for all} \quad x \in S.$$

   Then x^* is stable. (Note: as ξ_n are i.i.d., the condition does not depend on n.)

   Hint. Show that the process $V(x_n)$ is a supermartingale.

2. Prove the following theorem:

   Theorem A.2.2. *Suppose that there is a continuous function $V : S \to [0, \infty[$ with $V(x^*) = 0$ and $V(x) > 0$ for $x \neq x^*$, such that*

   $$\mathbb{E}(V(F(x, \xi_n))) - V(x) = k(x) < 0 \quad \textit{whenever} \quad x \neq x^*.$$

   Then x^* is globally stable.

   Hint. The proof proceeds roughly as follows. Fill in the steps:

   a) Write $V(x_0) - \mathbb{E}(V(x_n))$ as a telescoping sum. Use this and the condition in the theorem to prove that $k(x_n) \to 0$ in probability "fast enough".

   b) Prove that if some sequence $s_n \in S$ converges to a point $s \in S$, then $k(s_n) \to k(s)$, i.e., that $k(x)$ is a continuous function.

   c) As $k(x_n) \to 0$ a.s., k is continuous, and $k(s_n) \to 0$ only if $s_n \to x^*$ (why?), you can now conclude that $x_n \to x^*$ a.s.

3. **(Inverted pendulum in the rain)** A simple discrete time model for a controlled, randomly forced overdamped pendulum is

$$\theta_{n+1} = \theta_n + (1 + \xi_n)\sin(\theta_n)\Delta + u_{n+1}\Delta \qquad \mod 2\pi,$$

where θ_n is the angle ($\theta = 0$ is up) of the pendulum at time $n\Delta$, Δ is the time step size (be sure to take it small enough), u_{n+1} an applied control (using a servo motor), and ξ_n are i.i.d. random variables uniformly distributed on $[0, 1]$. The $\sin\theta_n$ term represents the downward gravitational force, while the term $\xi_n \sin\theta_n$ represents randomly applied additional forces in the downward direction—i.e., the force exerted on the pendulum by rain drops falling from above. *(This model is completely contrived! Don't take it too seriously.)*

Let us represent the circle $\theta \in S^1$ as the unit circle in $\mathbb{R}^2$. Writing $x_n = \sin\theta_n$, $y_n = \cos\theta_n$, and $f(x, \xi, u) = (1 + \xi)x\Delta + u\Delta$, we get

$$\begin{aligned} x_{n+1} &= x_n \cos(f(x_n, \xi_n, u_{n+1})) + y_n \sin(f(x_n, \xi_n, u_{n+1})), \\ y_{n+1} &= y_n \cos(f(x_n, \xi_n, u_{n+1})) - x_n \sin(f(x_n, \xi_n, u_{n+1})). \end{aligned}$$

Find some control law $u_{n+1} = g(x_n, y_n)$ that makes the inverted position $\theta = 0$ stable. (Try an intuitive control law and a linear Lyapunov function; you might want to use your favorite computer program to plot $k(\cdot)$.)

4. **Bonus question:** The previous results can be localized to a neighborhood. Prove the following modifications of the previous theorems:

Theorem A.2.3. *Suppose there is a continuous function $V : S \to [0, \infty[$ with $V(x^*) = 0$ and $V(x) > 0$ for $x \neq x^*$, and a neighborhood U of x^*, such that*

$$\mathbb{E}(V(F(x, \xi_n))) - V(x) = k(x) \leq 0 \quad \textit{whenever} \quad x \in U.$$

Then x^* is stable.

Theorem A.2.4. *Suppose there is a continuous function $V : S \to [0, \infty[$ with $V(x^*) = 0$ and $V(x) > 0$ for $x \neq x^*$, and a neighborhood U of x^*, such that*

$$\mathbb{E}(V(F(x, \xi_n))) - V(x) = k(x) < 0 \quad \textit{whenever} \quad x \in U \backslash \{x^*\}.$$

Then x^* is asymptotically stable.

Hint. Define a suitable stopping time τ, and apply the previous results to $x_{n\wedge\tau}$.

You can now show that the controlled pendulum is asymptotically stable.

A.3 Problem set 3

Q. 5. Let W_t be an n-dimensional Wiener process on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For non-random $x \in \mathbb{R}^n$, we call the process $W_t^x = x + W_t$ a Brownian motion *started at x*. We are going to investigate the behavior of this process in various dimensions.

1. Consider the annulus $D = \{x : r < \|x\| < R\}$ for some $0 < r < R < \infty$, and define the stopping time $\tau_x = \inf\{t : W_t^x \notin D\}$. For which functions $h : \mathbb{R}^n \to \mathbb{R}$ is $h(W^x_{t\wedge\tau_x})$ a martingale for all $x \in D$? You may assume that h is C^2 in some neighborhood of D. (Such functions are called *harmonic*).

2. Using the previous part, show that $h(x) = |x|$ is harmonic for $n = 1$, $h(x) = \log\|x\|$ is harmonic for $n = 2$, and $h(x) = \|x\|^{2-n}$ is harmonic for $n \geq 3$.

3. Let us write $\tau_x^R = \inf\{t : \|W_t^x\| \geq R\}$ and $\tau_x^r = \inf\{t : \|W_t^x\| \leq r\}$. What is $\mathbb{P}(\tau_x^r < \tau_x^R)$ for $n = 1, 2, 3, \ldots$? [**Hint**: $\|W^x_{\tau_x}\|$ can only take values r or R.]

4. What is $\mathbb{P}(\tau_x^r < \infty)$? Conclude the Brownian motion is *recurrent* for dimensions 1 and 2, but not for 3 and higher. [**Hint**: $\{\tau_x^r < \infty\} = \bigcup_{R>r}\{\tau_x^r < \tau_x^R\}$.]

Q. 6. We consider a single stock, which, if we were to invest one dollar at time zero, would be worth $S_t = e^{(\mu-\sigma^2/2)t+\sigma W_t}$ dollars by time t; here $\mu > 0$ (the return rate) and $\sigma > 0$ (the volatility) are constants, and W_t is a Wiener process on $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, \mathbb{P})$. We also have a bank account, which, if we were to deposit one dollar at time zero, would contain $R_t = e^{rt}$ dollars at time t, where $r > 0$ (the interest rate) is constant.

If we invest α_0 dollars in stock and β_0 dollars in the bank at time zero, then at time t our *total wealth* is $X_t = \alpha_0 S_t + \beta_0 R_t$ dollars. We can decide to reinvest at time t, so to put $\alpha_t S_t$ dollars in stock and $\beta_t R_t$ dollars in the bank. However, if our investment is *self-financing*, then we should make sure that $X_t = \alpha_0 S_t + \beta_0 R_t = \alpha_t S_t + \beta_t R_t$ (i.e., the total amount of invested money is the same: we have just transferred some money from stock to the bank or vice versa, without adding in any new money). Note that we will allow α_t and β_t to be negative: you can borrow money or sell short.

1. Show that if we modify our investment at times $t_1, t_2, \ldots$, then

$$X_{t_{n+1}} = \alpha_0 + \beta_0 + \sum_{i=0}^{n} \alpha_{t_i}(S_{t_{i+1}} - S_{t_i}) + \sum_{i=0}^{n} \beta_{t_i}(R_{t_{i+1}} - R_{t_i}),$$

   provided our strategy is self-financing. Show that this expression is identical to

$$X_{t_{n+1}} = X_0 + \int_0^{t_{n+1}} (\mu\alpha_s S_s + r\beta_s R_s)\, ds + \int_0^{t_{n+1}} \sigma\alpha_s S_s\, dW_s,$$

   where α_t and β_t are the simple integrands that take the values α_{t_i} and β_{t_i} on the interval $[t_i, t_{i+1}]$, respectively. [Assume that α_{t_i} and β_{t_i} are $\mathcal{F}_{t_i}$-measurable (obviously!) and sufficiently integrable.]

The integral expression for X_t still makes sense for continuous time strategies with $\alpha_t S_t$ and $\beta_t R_t$ in $\mathcal{L}^2(\mu_T \times \mathbb{P})$ (which we will always assume). Hence we can *define* a self-financing strategy to be a pair α_t, β_t that satisfies this expression (in addition to $X_t = \alpha_t S_t + \beta_t R_t$, of course). You can see this as a limit of discrete time strategies.

In a sensible model, we should not be able to find a reasonable strategy α_t, β_t that makes money for nothing. Of course, if we put all our money in the bank, then we

will always make money for sure just from the interest. It makes more sense to study the normalized market, where all the prices are *discounted* by the interest rate. So we will consider the discounted wealth $\overline{X}_t = X_t/R_t$ and stock price $\overline{S}_t = S_t/R_t$. We want to show that there does not exist a trading strategy with $\overline{X}_0 = a$, $\overline{X}_t \geq a$ a.s., and $\mathbb{P}(\overline{X}_t > a) > 0$. Such a money-for-nothing opportunity is called *arbitrage*.

2. Show that the discounted wealth at time t is given by

$$\overline{X}_t = X_0 + \int_0^t (\mu - r)\alpha_s \overline{S}_s \, ds + \int_0^t \sigma \alpha_s \overline{S}_s \, dW_s.$$

3. Find a new measure $\mathbb{Q}$ such that $\mathbb{Q} \ll \mathbb{P}$, $\mathbb{P} \ll \mathbb{Q}$, and $\overline{X}_t$ is a martingale under $\mathbb{Q}$ (for reasonable α_t). $\mathbb{Q}$ is called the *equivalent martingale measure*.

4. The equivalent martingale measure has a very special property: $\mathbb{E}_{\mathbb{Q}}(\overline{X}_t) = X_0$ (assuming our initial wealth X_0 is non-random), regardless of the trading strategy. Use this to prove that there is no arbitrage in our model.

We are going to do some simple option pricing theory. Consider something called a *European call option*. This is a contract that says the following: at some predetermined time T (the *maturity*), we are allowed to buy one unit of stock at some predetermined price K (the *strike price*). This is a sort of insurance against the stock price going very high: if the stock price goes below K by time T we can still buy stock at the market price, and we only lose the money we paid to take out the option; if the stock price goes above K by time T, then we make money as we can buy the stock below the market price. The total payoff for us is thus $(S_T - K)^+$, minus the option price. The question is what the seller of the option should charge for that service.

5. If we took out the option, we would make $(S_T - K)^+$ dollars (excluding the option price). Argue that we could obtain exactly the same payoff by implementing a particular trading strategy α_t, β_t, a *hedging strategy*, provided that we have sufficient starting capital (i.e., for some X_0, α_t, β_t, we actually have $X_T = (S_T - K)^+$). Moreover, show that there is only one such strategy.

6. Argue that the starting capital required for the hedging strategy is the only fair price for the option. (If a different price is charged, either we or the seller of the option can make money for nothing.)

7. What is the price of the option? [**Hint:** use the equivalent martingale measure.]

Congratulations—you have just developed the famous Black-Scholes model!

Q. 7 (Bonus question: baby steps in the Malliavin calculus). *Very roughly speaking*, whereas the Itô calculus defines integrals $\int \cdots dW_t$ with respect to the Wiener process, the Malliavin calculus defines *derivatives* "$d\cdots/dW_t$" with respect to the Wiener process. This has applications both in stochastic analysis (smoothness of densities, anticipative calculus) and in finance (computation of sensitivities and hedging strategies, variance reduction of Monte Carlo simulation, insider trading models, etc.)

This is a much more advanced topic than we are going deal with in this course. As we have the necessary tools to get started, however, I can't resist having you explore some of the simplest ideas (for fun and extra credit—this is not a required problem!).

We work on $(\Omega, \mathcal{F}, \mathbb{P})$, on which is defined a Wiener process W_t with its natural filtration $\mathcal{F}_t = \sigma\{W_s : s \le t\}$. We restrict ourselves to a finite time interval $t \in [0, T]$. An $\mathcal{F}_T$-measurable random variable X is called *cylindrical* if it can be written as $X = f(W_{t_1}, \dots, W_{t_n})$ for a finite number of times $0 < t_1 < \cdots < t_n \le T$ and some function $f \in C_0^\infty$. For such X, the *Malliavin derivative* of X is defined as

$$\mathbf{D}_t X = \sum_{i=1}^{n} \frac{\partial f}{\partial x^i}(W_{t_1}, \dots, W_{t_n})\, I_{t \le t_i}.$$

1. For cylindrical X, prove the *Clark-Ocone formula:*

$$X = \mathbb{E}(X) + \int_0^T \mathbb{E}(\mathbf{D}_t X | \mathcal{F}_t)\, dW_t.$$

Hint: look at the proofs of lemma 4.6.5 and lemma 3.1.9.

As any $\mathcal{F}_T$-measurable random variable Y in $\mathcal{L}^2(\mathbb{P})$ can be approximated by cylindrical functions, one can now extend the definition of the Malliavin derivative to a much larger class of random variables by taking limits. Not all such Y are Malliavin differentiable, but with a little work one can define a suitable Sobolev space of differentiable random variables. If you want to learn more about this, see [Nua95].

Let us take a less general approach (along the lines of Clark's original result), which allows a beautiful alternative development of the Clark-Ocone formula (the idea is due to Haussmann and Bismut, here we follow D. Williams). Let $f : C([0, T]) \to \mathbb{R}$ be a measurable map. We will consider random variables of the form $X = f(W_\cdot)$ (actually, any $\mathcal{F}_T$-measurable random variable can be written in this way.)

2. Let u_t be bounded and $\mathcal{F}_t$-adapted, and let $\varepsilon \in \mathbb{R}$. Prove the *invariance formula*

$$\mathbb{E}(f(W_\cdot)) = \mathbb{E}\left[f\left(W_\cdot - \varepsilon \int_0^\cdot u_s\, ds \right) e^{\varepsilon \int_0^T u_s\, dW_s - \frac{\varepsilon^2}{2} \int_0^T (u_s)^2 ds} \right].$$

We are now going to impose a (Fréchet) differentiability condition on f. We assume that for any continuous function x and bounded function α on $[0, T]$, we have

$$f\left(x_\cdot + \varepsilon \int_0^\cdot \alpha_s\, ds \right) - f(x_\cdot) = \varepsilon \int_0^T f'(s, x_\cdot)\, \alpha_s\, ds + o(\varepsilon),$$

where $f' : [0, T] \times C([0, T]) \to \mathbb{R}$ is some measurable function. Then for $X = f(W_\cdot)$, we define the Malliavin derivative of X as $\mathbf{D}_t X = f'(t, W_\cdot)$.

3. Show that this definition of $\mathbf{D}_t X$ coincides with our previous definition for cylindrical random variables X.

4. Let $X = f(W_\cdot)$, and assume for simplicity that $f(x_\cdot)$ and $f'(t, x_\cdot)$ are bounded. By taking the derivative with respect to ε, at $\varepsilon = 0$, of the invariance formula above, prove the *Malliavin integration by parts formula*

$$\mathbb{E}\left[X \int_0^T u_s \, dW_s\right] = \mathbb{E}\left[\int_0^T u_s \, \mathbf{D}_s X \, ds\right]$$

for any bounded and $\mathcal{F}_t$-adapted process u_t. Show, furthermore, that

$$\mathbb{E}\left[X \int_0^T u_s \, dW_s\right] = \mathbb{E}\left[\int_0^T u_s \, \mathbb{E}(\mathbf{D}_s X | \mathcal{F}_s) \, ds\right].$$

5. Using the Itô representation theorem, prove that there is a unique $\mathcal{F}_t$-adapted process C_t such that for any bounded and $\mathcal{F}_t$-adapted process u_t

$$\mathbb{E}\left[X \int_0^T u_s \, dW_s\right] = \mathbb{E}\left[\int_0^T u_s \, C_s \, ds\right].$$

Conclude that the Clark-Ocone formula still holds in this context.

A.4 Problem set 4

This problem set involves some programming; you may use whatever you want for this, but I strongly recommend you use either `Matlab` (or something similar, such as `R`) or a compiled programming language (e.g., `C++`) for this purpose. If you have never done any programming, please contact me and we will figure something out.

Q. 8. Consider the stochastic differential equations

$$dX_t^r = \sin(X_t^r) \, dW_t^1 + \cos(X_t^r) \, dW_t^2, \quad X_0^r = r, \qquad dY_t^r = dW_t^1, \quad Y_0^r = r,$$

where $r \in \mathbb{R}$ is non-random and (W_t^1, W_t^2) is a two-dimensional Wiener process.

1. Show that X_t^r has the same law as Y_t^r for every fixed time t.

   [**Hint:** investigate the Kolmogorov backward equations for X_t^r and Y_t^r.]

2. Show that X_t^r has independent increments. Together with the previous part, this implies that $\{X_t^r\}$ is a one-dimensional Brownian motion started at r.

   [**Hint:** show that $\mathbb{E}(f(X_t^r - X_s^r)|\mathcal{F}_s) = \mathbb{E}(f(X_t^r - z)|\mathcal{F}_s)|_{z=X_s^r} \equiv g(X_s^r)$ is constant, i.e., the function $g(x)$ is independent of x (you do not need to prove the first equality; it follows as in the proof of lemma 3.1.9). Then show why this implies $\mathbb{E}(f(X_t^r - X_s^r)Z) = \mathbb{E}(f(X_t^r - X_s^r)) \, \mathbb{E}(Z)$ for any $\mathcal{F}_s$-measurable Z.]

X_t^r is thus a Brownian motion started at r—what more can be said? Surprisingly, X_t^r and Y_t^r behave very differently if we consider multiple initial points $r_1, \ldots, r_n$ simultaneously, *but driven by the same noise*. In other words, we are interested in

$$Y_t = (Y_t^{r_1}, \ldots, Y_t^{r_n}) = (r_1 + W_t^1, \ldots, r_n + W_t^1), \qquad X_t = (X_t^{r_1}, \ldots, X_t^{r_n}),$$

where the latter is the solution of the n-dimensional SDE every component of which satisfies the equation for X_t^r above.

3. Use the Euler-Maruyama method to compute several sample paths of X_t and of Y_t in the interval $t \in [0, 10]$, with $(r_1, \dots, r_n) = (-3, -2.5, -2 \dots, 3)$ and with step size $\Delta t = .001$. Qualitatively, what do you see?

Apparently the SDEs for X_t^r and Y_t^r are qualitatively different, despite that for every initial condition their solutions have precisely the same law! These SDEs generate the same Markov process, but a different *flow* $r \mapsto X_t^r$, $r \mapsto Y_t^r$. Stochastic flows are important in random dynamics (they can be used to define Lyapunov exponents, etc.), and have applications, e.g., in the modelling of ocean currents.

Q. 9. We are going to investigate the inverted pendulum of example 6.6.5, but with a different cost functional. Recall that we set

$$d\theta_t^u = c_1 \sin(\theta_t^u)\, dt - c_2 \cos(\theta_t^u)\, u_t\, dt + \sigma\, dW_t.$$

As the coefficients of this equation are periodic in θ, we may interpret its solution modulo 2π (i.e., θ_t^u evolves on the circle, which is of course the intention).

Our goal is to keep θ_t^u as close to the up position $\theta = 0$ as possible on some reasonable time scale. We will thus investigate the discounted cost

$$J_\lambda[u] = \mathbb{E}\left[\int_0^\infty e^{-\lambda s}\{p\,(u_s)^2 + q\,(1 - \cos(\theta_s^u))\}\, ds\right].$$

This problem does not lend itself to analytic solution, so we approach it numerically.

1. Starting from the appropriate Bellman equation, develop a Markov chain approximation to the control problem of minimizing $J_\lambda[u]$ following the finite-difference approach of section 6.6. Take the fact that θ_t^u evolves on the circle into account to introduce appropriate boundary conditions.

   [**Hint:** it is helpful to realize what the discrete dynamic programming equation for a discounted cost looks like. If x_n^α is a controlled Markov chain with transition probabilies $P_{i,j}^\alpha$ from state i to state j under the control α, and

   $$K_\varrho[u] = \mathbb{E}\left[\sum_{n=0}^\infty \varrho^n\, w(x_n^u, u_{n+1})\right], \qquad 0 < \varrho < 1,$$

   then the value function satisfies $V(i) = \min_{\alpha \in \mathbb{U}}\{\varrho \sum_j P_{i,j}^\alpha V(j) + w(i, \alpha)\}$. You will prove a verification theorem for such a setting in part 2.]

2. To which discrete optimal control problem does your numerical method correspond? Prove an analog of proposition 6.6.2 for this case.

3. Using the Jacobi iteration method, implement the numerical scheme you developed, and plot the optimal control and the value function.

   You can try, for example, $c_1 = c_2 = \sigma = .5$, $p = q = 1$, $\lambda = .1$; a grid which divides $[0, 2\pi[$ into 100 points; and 500 iterations of the Jacobi method (but play around with the parameters and see what happens, if you are curious!)

A.5 Problem set 5

Q. 10. A beautiful butterfly is fluttering around a patch of tasty flowers. At a certain time, the butterfly decides that it has got the most out of its current flower patch, and flies off at a rapid rate in search of fresh flowers. We model the position x_t of the butterfly at time t (in one dimension for simplicity, i.e., $x_t \in \mathbb{R}^1$) by the equation

$$dx_t = \gamma\, I_{\tau \le t}\, dt + \sigma\, dB_t, \qquad x_0 = x,$$

where σ determines the vigorousness of the butterfly's fluttering, τ is the time at which it decides to fly away, γ is the speed at which it flies away, and B_t is a Wiener process. We will assume that τ is exponentially distributed, i.e., that $\mathbb{P}(\tau > t) = e^{-\lambda t}$.

Beside the butterfly the forest also features a biologist, who has come equipped with a butterfly net and a Segway. The biologist can move around at will on his Segway by applying some amount of power u_t; his position z_t^u is then given by

$$\frac{dz_t^u}{dt} = \beta\, u_t, \qquad z_0 = z.$$

Mesmerized by the colorful butterfly, the biologist hatches a plan: he will try to intercept the butterfly at a fixed time T, so that he can catch it and bring it back to his laboratory for further study. However, he would like to keep his total energy consumption low, because he knows from experience that if he runs the battery in the Segway dry he will flop over (and miss the butterfly). As such, the biologist wishes to pursue the butterfly using a strategy u that minimizes the cost functional

$$J[u] = \mathbb{E}\left[P \int_0^T (u_t)^2\, dt + Q\,(x_T - z_T^u)^2 \right], \qquad P, Q > 0,$$

where the first term quantifies the total energy consumption and the second term quantifies the effectiveness of the pursuit. The entire setup is depicted in figure A.1.

Note that this is a partially observed control problem: the control u_t is allowed to be $\mathcal{F}_t^x = \sigma\{x_s : s \le t\}$-adapted, as the biologist can *see* where the butterfly is, but the biologist does not know the time τ at which the butterfly decides to leave.

1. Define the *predicted interception point* $r_t = \mathbb{E}(x_T | \mathcal{F}_t^x)$. Show that

$$r_t = x_t + \gamma \int_t^T \mathbb{P}(\tau \le s | \mathcal{F}_t^x)\, ds.$$

2. Prove that for $s > t$, we have $1 - \mathbb{P}(\tau \le s | \mathcal{F}_t^x) = e^{-\lambda(s-t)}(1 - \mathbb{P}(\tau \le t | \mathcal{F}_t^x))$. Now obtain an explicit expression for r_t in terms of x_t and $\pi_t = \mathbb{P}(\tau \le t | \mathcal{F}_t^x)$.

3. Using Itô's rule and the appropriate filter, find a stochastic differential equation for (r_t, π_t) which is driven by the innovations process $\bar{B}_t$ and in which τ no longer appears explicitly.

Figure A.1. Schematic of problem 10 ("Der Schmetterlingsjäger", Carl Spitzweg, 1840).

4. Define $e_t^u = r_t - z_t^u$. Obtain a stochastic differential equation for (e_t^u, π_t) which is driven by $\bar{B}_t$, and rewrite the cost $J[u]$ in terms of (e_t^u, π_t). You have now converted the partially observed control problem into one with complete observations. What is the corresponding Bellman equation?

5. We are now faced with the difficulty of solving a nonlinear control problem, but nonetheless we will find an analytic solution for the optimal control. To this end, try substituting into the Bellman equation a value function of the form $V(t, e, \pi) = a(t)\, e^2 + b(t, \pi)$, and find equations for $a(t)$ and $b(t, \pi)$. Use this to determine the optimal control strategy. You may assume that the equation you find for $b(t, \pi)$ admits a sufficiently smooth solution (this is in fact the case).

6. Roughly speaking, how could you interpret the optimal strategy? (This is not a deep question, give a one or two line answer.)

A more general version of this problem can be found in [HR92].

Bibliography

[Alv04] L. H. R. Alvarez, *Stochastic forest stand value and optimal timber harvesting*, SIAM J. Control Optim. **42** (2004), 1972–1993.

[Apo69] T. M. Apostol, *Calculus. Volume II*, Wiley, 1969.

[Arn74] L. Arnold, *Stochastic differential equations: Theory and applications*, Wiley, 1974.

[Bac00] L. Bachelier, *Théorie de la spéculation*, Ann. Sci. E.N.S. Sér. 3 **17** (1900), 21–86.

[BD98] M. Boué and P. Dupuis, *A variational representation for certain functionals of Brownian motion*, Ann. Probab. **26** (1998), 1641–1659.

[Ben82] A. Bensoussan, *Stochastic control by functional analysis methods*, North Holland, 1982.

[Ben92] ———, *Stochastic control of partially observable systems*, Cambridge University Press, 1992.

[BHL99] D. Brigo, B. Hanzon, and F. Le Gland, *Approximate nonlinear filtering by projection on exponential manifolds of densities*, Bernoulli **5** (1999), 495–534.

[Bic02] K. Bichteler, *Stochastic integration with jumps*, Cambridge University Press, 2002.

[Bil86] P. Billingsley, *Probability and measure*, second ed., Wiley, 1986.

[Bil99] ———, *Convergence of probability measures*, second ed., Wiley, 1999.

[Bir67] G. Birkhoff, *Lattice theory*, third ed., AMS, 1967.

[Bis81] J.-M. Bismut, *Mécanique aléatoire*, Lecture Notes in Mathematics 866, Springer, 1981.

[BJ87] R. S. Bucy and P. D. Joseph, *Filtering for stochastic processes with applications to guidance*, second ed., Chelsea, 1987.

[BK29] S. Banach and C. Kuratowski, *Sur une généralisation du problème de la mesure*, Fund. Math. **14** (1929), 127–131.

[BL82] A. Bensoussan and J. L. Lions, *Application of variational inequalities in stochastic control*, North Holland, 1982.

[BL84] ———, *Impulse control and quasi-variational inequalities*, Gauthier-Villars, 1984.

[BM04] A. J. Berglund and H. Mabuchi, *Feedback controller design for tracking a single fluorescent molecule*, Appl. Phys. B **78** (2004), 653–659.

[BN93] M. Basseville and I. V. Nikiforov, *Detection of abrupt changes. theory and application*, Prentice Hall, available at `www.irisa.fr/sisthem/kniga/`, 1993.

[BØ94] K. A. Brekke and B. Øksendal, *Optimal switching in an economic activity under uncertainty*, SIAM J. Control Optim. **32** (1994), 1021–1036.

[Bor00] C. Borell, *Diffusion equations and geometric inequalities*, Potential Anal. **12** (2000), 49–71.

[Bor05] V. S. Borkar, *Controlled diffusion processes*, Probab. Surv. **2** (2005), 213–244.

[BR05] A. Beskos and G. O. Roberts, *Exact simulation of diffusions*, Ann. Appl. Probab. **15** (2005), 2422–2444.

[Bro28] R. Brown, *A brief account of microscopical observations made in the months of June, July and August, 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies*, Phil. Mag. **4** (1828), 161–173.

[BS73] F. Black and M. Scholes, *The pricing of options and corporate liabilities*, J. Polit. Econ. **81** (1973), 637–654.

[Chi04] P. Chigansky, *Introduction to nonlinear filtering*, lecture notes, available at `www.wisdom.weizmann.ac.il/~pavel/`, 2004.

[CL97] D. Crisan and T. Lyons, *Nonlinear filtering and measure-valued processes*, Probab. Th. Rel. Fields **109** (1997), 217–244.

[Cri02] D. Crisan, *Numerical methods for solving the stochastic filtering problem*, Numerical methods and stochastics (Toronto, ON, 1999), Fields Inst. Commun., vol. 34, AMS, 2002, pp. 1–20.

[Dav77] M. H. A. Davis, *Linear estimation and stochastic control*, Chapman and Hall, 1977.

[Dav79] ———, *Martingale methods in stochastic control*, Stochastic control theory and stochastic differential systems, Lecture Notes in Control and Information Sci., vol. 16, Springer, 1979, pp. 85–117.

[Dav80] ———, *The representation of functionals of diffusion processes as stochastic integrals*, Trans. Cam. Phil. Soc. **87** (1980), 157–166.

[DE97] P. Dupuis and R. S. Ellis, *A weak convergence approach to the theory of large deviations*, Wiley, 1997.

[DE06] M. H. A. Davis and A. Etheridge, *Louis Bachelier's theory of speculation: The origins of modern finance*, Princeton University Press, 2006.

[Del04] P. Del Moral, *Feynman-Kac formulae*, Springer, 2004, Genealogical and interacting particle systems with applications.

[DK03] S. Dayanik and I. Karatzas, *On the optimal stopping problem for one-dimensional diffusions*, Stochastic Process. Appl. **107** (2003), 173–212.

[DKW01] H. Deng, M. Krstić, and R. J. Williams, *Stabilization of stochastic nonlinear systems driven by noise of unknown covariance*, IEEE Trans. Automat. Control **46** (2001), 1237–1253.

[DM78] C. Dellacherie and P.-A. Meyer, *Probabilities and potential*, North-Holland, 1978.

[DM82] ———, *Probabilities and potential. B*, North-Holland, 1982.

[DO94] P. Dupuis and J. Oliensis, *An optimal control formulation and related numerical methods for a problem in shape reconstruction*, Ann. Appl. Probab. **4** (1994), 287–346.

[Doo53] J. L. Doob, *Stochastic processes*, Wiley, 1953.

[Dud02] R. M. Dudley, *Real analysis and probability*, Cambridge University Press, 2002.

[Duf01] D. Duffie, *Dynamic asset pricing theory*, third ed., Princeton University Press, 2001.

[Dyn06] E. B. Dynkin, *Theory of Markov processes*, Dover, 2006.

[EAM95] R. J. Elliott, L. Aggoun, and J. B. Moore, *Hidden Markov models*, Springer, 1995.

[Ein05] A. Einstein, *Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen*, Ann. Phys. **17** (1905), 549–560.

[EK86] S. N. Ethier and T. G. Kurtz, *Markov processes: Characterization and convergence*, Wiley, 1986.

[Ell82] R. J. Elliott, *Stochastic calculus and applications*, Springer, 1982.

[FM83] W. H. Fleming and S. K. Mitter, *Optimal control and nonlinear filtering for nondegenerate diffusion processes*, Stochastics **8** (1982/83), 63–77.

[FR75] W. H. Fleming and R. W. Rishel, *Deterministic and stochastic optimal control*, Springer, 1975.

[Fri75] A. Friedman, *Stochastic differential equations and applications*, Academic Press, 1975.

[FS06] W. H. Fleming and H. M. Soner, *Controlled Markov processes and viscosity solutions*, second ed., Springer, 2006.

[FW98] M. I. Freidlin and A. D. Wentzell, *Random perturbations of dynamical systems*, second ed., Springer, 1998.

[GP84] A. Germani and M. Piccioni, *A Galerkin approximation for the Zakai equation*, System modelling and optimization (Copenhagen, 1983), Lecture Notes in Control and Inform. Sci., vol. 59, Springer, 1984, pp. 415–423.

[GS96] I. I. Gikhman and A. V. Skorokhod, *Introduction to the theory of random processes*, Dover, 1996.

[GS01] G. R. Grimmett and D. R. Stirzaker, *Probability and random processes*, third ed., Oxford University Press, 2001.

[Han07] F. B. Hanson, *Applied stochastic processes and control for jump-diffusions: Modeling, analysis and computation*, SIAM, to appear, draft available at `www2.math.uic.edu/~hanson/`, 2007.

[Has80] R. Z. Has'minskiĭ, *Stochastic stability of differential equations*, Sijthoff & Noordhoff, 1980.

[HC99] J. L. Hibey and C. D. Charalambous, *Conditional densities for continuous-time nonlinear hybrid systems with applications to fault detection*, IEEE Trans. Automat. Control **44** (1999), 2164–2169.

[Hid80] T. Hida, *Brownian motion*, Springer, 1980.

[HØUZ96] H. Holden, B. Øksendal, J. Ubøe, and T. Zhang, *Stochastic partial differential equations. A modeling, white noise functional approach*, Birkhäuser, 1996.

[HR92] K. Helmes and R. W. Rishel, *The solution of a partially observed stochastic optimal control problem in terms of predicted miss*, IEEE Trans. Automat. Control **37** (1992), 1462–1464.

[HW81] M. Hazewinkel and J. C. Willems (eds.), *Stochastic systems: the mathematics of filtering and identification and applications*, NATO Advanced Study Institute Series C: Mathematical and Physical Sciences, vol. 78, D. Reidel, 1981.

[Itô44] K. Itô, *Stochastic integral*, Proc. Imp. Acad. Tokyo **20** (1944), 519–524.

[IW89] N. Ikeda and S. Watanabe, *Stochastic differential equations and diffusion processes*, second ed., North-Holland, 1989.

[JZ06] A. Jack and M. Zervos, *Impulse control of one-dimensional Ito diffusions with an expected and a pathwise ergodic criterion*, Appl. Math. Opt. **54** (2006), 71–93.

[Kal80] G. Kallianpur, *Stochastic filtering theory*, Springer, 1980.

[Kal97] O. Kallenberg, *Foundations of modern probability*, Springer, 1997.

[Kar03] I. Karatzas, *A note on Bayesian detection of change-points with an expected miss criterion*, Statist. Decisions **21** (2003), 3–14.

[KD01] H. J. Kushner and P. Dupuis, *Numerical methods for stochastic control problems in continuous time*, second ed., Springer, 2001.

[Kha02] H. K. Khalil, *Nonlinear systems*, third ed., Prentice Hall, 2002.

[Kor99] R. Korn, *Some applications of impulse control in mathematical finance*, Math. Meth. Oper. Res. **50** (1999), 493–518.

[KP92] P. E. Kloeden and E. Platen, *Numerical solution of stochastic differential equations*, Springer, 1992.

[Kri05] V. Krishnan, *Nonlinear filtering and smoothing: An introduction to martingales, stochastic integrals and estimation*, Dover, 2005.

[Kry80] N. V. Krylov, *Controlled diffusion processes*, Springer, 1980.

[KS72] H. Kwakernaak and R. Sivan, *Linear optimal control systems*, Wiley, 1972, avaliable at `www.ieeecss.org/PAB/classics/`.

[KS91] I. Karatzas and S. E. Shreve, *Brownian motion and stochastic calculus*, second ed., Springer, 1991.

[KS98] ———, *Methods of mathematical finance*, Springer, 1998.

[Kun84] H. Kunita, *Stochastic differential equations and stochastic flows of diffeomorphisms*, École d'Été de Probabilités de Saint-Flour XII (P. L. Hennequin, ed.), Lecture Notes in Mathematics 1097, Springer, 1984.

[Kun90] ———, *Stochastic flows and stochastic differential equations*, Cambridge University Press, 1990.

[Kus67] H. J. Kushner, *Stochastic stability and control*, Academic Press, 1967.

[Kus71] ———, *Introduction to stochastic control*, Holt, Rinehart and Winston, 1971.

[Kus72] ———, *Stochastic stability*, Stability of Stochastic Dynamical systems (R.F. Curtain, ed.), Lecture Notes in Mathematics, vol. 294, Springer, 1972, pp. 97–123.

[Kus77] ———, *Probability methods for approximations in stochastic control and for elliptic equations*, Academic Press, 1977.

[Kus84] ———, *Approximation and weak convergence methods for random processes, with applications to stochastic systems theory*, MIT Press, 1984.

[Kus90] ———, *Numerical methods for stochastic control problems in continuous time*, SIAM J. Control Optim. **28** (1990), 999–1048.

[KV86] P. R. Kumar and P. Varaiya, *Stochastic systems: Estimation, identification and adaptive control*, Prentice Hall, 1986.

[Lar05] S. Larsson, *Numerical methods for stochastic ODEs*, lecture notes, available at `www.math.chalmers.se/~stig/`, 2005.

[Let88] G. Letta, *Un exemple de processus mesurable adapté non-progressif*, Sém. Probab. Strasbourg **22** (1988), 449–453.

[LL01] E. H. Lieb and M. Loss, *Analysis*, second ed., American Mathematical Society, 2001.

[LMR97] S. Lototsky, R. Mikulevicius, and B. L. Rozovskii, *Nonlinear filtering revisited: a spectral approach*, SIAM J. Control Optim. **35** (1997), 435–461.

[LS01a] R. S. Liptser and A. N. Shiryaev, *Statistics of random processes I. General theory*, second ed., Springer, 2001.

[LS01b] ———, *Statistics of random processes II. Applications*, second ed., Springer, 2001.

[Mak86] A. M. Makowski, *Filtering formulae for partially observed linear systems with non-Gaussian initial conditions*, Stochastics **16** (1986), 1–24.

[McK69] H. P. McKean, *Stochastic integrals*, Academic Press, 1969.

[Mer71] R. C. Merton, *Optimum consumption and portfolio rules in a continuous-time model*, J. Econ. Th. **3** (1971), 373–413.

[Mit82] S. K. Mitter, *Lectures on nonlinear filtering and stochastic control*, Nonlinear filtering and stochastic control (Cortona, 1981), Lecture Notes in Math., vol. 972, Springer, 1982, pp. 170–207.

[MN03] S. K. Mitter and N. J. Newton, *A variational approach to nonlinear estimation*, SIAM J. Control Optim. **42** (2003), 1813–1833.

[MP06] P. Mörters and Y. Peres, *Brownian motion*, draft book, available at `www.stat.berkeley.edu/users/peres/bmbook.pdf`, 2006.

[Nev75] J. Neveu, *Discrete-parameter martingales*, North-Holland, 1975.

[Nua95] D. Nualart, *The Malliavin calculus and related topics*, Springer, 1995.

[Øks03] B. Øksendal, *Stochastic differential equations*, sixth ed., Springer, 2003.

[ØS05] B. Øksendal and A. Sulem, *Applied stochastic control of jump diffusions*, Springer, 2005.

[Par91] É. Pardoux, *Filtrage non linéaire et équations aux dérivées partielles stochastiques associées*, École d'Été de Probabilités de Saint-Flour XIX—1989, Lecture Notes in Math., vol. 1464, Springer, 1991, pp. 67–163.

[Par82] ———, *Équations du filtrage non linéaire, de la prédiction et du lissage*, Stochastics **6** (1981/82), 193–231.

[Pet06] I. R. Petersen, *Minimax LQG control*, Int. J. Appl. Math. Comput. Sci. **16** (2006), 309–323.

[Pol02] D. Pollard, *A user's guide to measure theoretic probability*, Cambridge University Press, 2002.

[Pro04] P. Protter, *Stochastic integration and differential equations*, second ed., Springer, 2004.

[PS93] P. Protter and J. San Martín, *General change of variable formulas for semimartingales in one and finite dimensions*, Probab. Th. Rel. Fields **97** (1993), 363–381.

[PS06] G. Peskir and A. N. Shiryaev, *Optimal stopping and free-boundary problems*, Birkhäuser, 2006.

[Rao72] M. Rao, *On modification theorems*, Trans. AMS **167** (1972), 443–450.

[RH06] R. Rishel and K. Helmes, *A variational inequality sufficient condition for optimal stopping with application to an optimal stock selling problem*, SIAM J. Control Optim. **45** (2006), 580–598.

[Rob83] M. Robin, *Long-term average cost control problems for continuous time Markov processes: a survey*, Acta Appl. Math. **1** (1983), 281–299.

[Roz90] B. L. Rozovskiĭ, *Stochastic evolution systems*, Kluwer, 1990.

[RS80] M. Reed and B. Simon, *Functional analysis*, second ed., Academic Press, 1980.

[RW00a] L. C. G. Rogers and D. Williams, *Diffusions, Markov processes, and martingales. Vol. 1*, Cambridge University Press, 2000.

[RW00b] ———, *Diffusions, Markov processes, and martingales. Vol. 2*, Cambridge University Press, 2000.

[RY99] D. Revuz and M. Yor, *Continuous martingales and Brownian motion*, third ed., Springer, 1999.

[Seg77] A. Segall, *Optimal control of noisy finite-state Markov processes*, IEEE Trans. Automat. Control **AC-22** (1977), 179–186.

[She91] S. J. Sheu, *Some estimates of the transition density of a nondegenerate diffusion Markov process*, Ann. Probab. **19** (1991), 538–561.

[Shi63] A. N. Shiryaev, *On optimum methods in quickest detection problems*, Theory Probab. Appl. **8** (1963), 22–46.

[Shi73] ———, *Statistical sequential analysis*, AMS, 1973.

[SM79] J. Szpirglas and G. Mazziotto, *Théorème de séparation dans le problème d'arrêt optimal*, Sém. Probab. Strasbourg **13** (1979), 378–384.

[Ste01] J. M. Steele, *Stochastic calculus and financial applications*, Springer, 2001.

[SV72] D. W. Stroock and S. R. Varadhan, *On the support of diffusion processes with applications to the strong maximum principle*, Proc. 6th Berkely Sympos. Math. Statist Prob., vol. III, 1972, pp. 333–368.

[Twa96] K. Twardowska, *Wong-Zakai approximations for stochastic differential equations*, Acta Appl. Math. **43** (1996), 317–359.

[Wie23] N. Wiener, *Differential space*, J. Math. Phys. **2** (1923), 131–174.

[Wil74] A. S. Willsky, *Fourier series and estimation on the circle with applications to synchronous communication—Part I: Analysis*, IEEE Trans. Inf. Th. **IT-20** (1974), 577–583.

[Wil91] D. Williams, *Probability with martingales*, Cambridge University Press, 1991.

[Wil98] Y. Willassen, *The stochastic rotation problem: a generalization of Faustmann's formula to stochastic forest growth*, J. Econ. Dyn. Control **22** (1998), 573–596.

[Won65] W. M. Wonham, *Some applications of stochastic differential equations to optimal nonlinear filtering*, SIAM J. Control **2** (1965), 347–369.

[Won68a] ———, *On a matrix Riccati equation of stochastic control*, SIAM J. Control **6** (1968), 681–697.

[Won68b] ———, *On the separation theorem of stochastic control*, SIAM J. Control **6** (1968), 312–326.

[WZ65] E. Wong and M. Zakai, *On the convergence of ordinary integrals to stochastic integrals*, Ann. Math. Stat. **36** (1965), 1560–1564.

[YZ99] J. Yong and X. Y. Zhou, *Stochastic controls*, Springer, 1999.