Stochastic Differential Equations: A Systems-Theoretic Approach (DRAFT)

Maxim Raginsky

May 1, 2023

Contents

Part I

Preliminaries

Chapter 1

Introduction

To Be Written

1.1 Notes and further reading

Mortensen [Mor69] gives a good discussion of the engineering modeling aspects of stochastic calculus.

Chapter 2

Brownian Motion and Diffusion Processes

2.1 Brownian motion

2.1.1 Definition and construction

The standard one-dimensional Brownian motion (also known as the Wiener process) is a random process $W = (W_t)_{t\geq 0}$ with the following properties:

1. $W_0 = 0$, and the increments $W_{t_1}, W_{t_2} - W_{t_1}, W_{t_3} - W_{t_2}, \ldots$ for $0 < t_1 < t_2 < t_3 < \ldots$ are independent.

2. $W_t - W_s$ is Gaussian with zero mean and variance $|t - s|$.

3. The sample paths $t \mapsto W_t$ are continuous almost surely.

The first two items imply that W is a zero-mean Gaussian process with covariance function $\mathbf{E}[W_t W_s] = t \wedge s$. There are a number of ways one could construct a Brownian motion. We will illustrate one that has the advantage of being fairly direct and elementary[1] in the sense of requiring the minimum of fancy probabilistic machinery.

To start with, we note that it suffices to construct the process only on the unit interval, i.e., for $t \in [0, 1]$. Once this is done, we can generate an infinite sequence of independent copies of W on $[0, 1]$, i.e., $W^{(0)}, W^{(1)}, W^{(2)}, \ldots$ and then take

$$W_t = \begin{cases} W_t^{(0)}, & 0 \leq t < 1 \\ W_1^{(0)} + W_{t-1}^{(1)}, & 1 \leq t < 2 \\ W_1^{(0)} + W_2^{(1)} + W_{t-2}^{(2)}, & 2 \leq t < 3 \\ \ldots \end{cases} \tag{2.1.1}$$

– indeed, it is not hard to verify that the resulting process has the properties 1–3 above. We can now motivate the construction of $W := (W_t)_{t\in[0,1]}$ formally as follows: Let $\xi = (\xi_t)_{t\in[0,1]}$ be a white noise

[1] "Elementary" here does not mean "simple;" it just means that no advanced techniques are necessary beyond basic properties of Gaussian randomm variables and some real analysis.

process, i.e., the time derivative of W: $\xi_t = \dot{W}_t$. Let us pick a complete orthonormal basis $(\psi_n)_{n\in\mathbb{Z}_+}$ of the Hilbert space $L^2[0,1]$ and expand ξ in this basis:

$$\xi_t = \sum_{n=0}^{\infty} Z_n \psi_n(t), \qquad t \in [0,1]. \tag{2.1.2}$$

This is a (formal) Karhunen–Loève expansion of the white noise. Here, the coefficients Z_n are *random* and given by

$$Z_n = \langle \xi, \psi_n \rangle = \int_0^1 \xi_t \psi_n(t)\,\mathrm{d}t. \tag{2.1.3}$$

Evidently, $\mathbf{E}[Z_n] = 0$ and, since $\mathbf{E}[\xi_t \xi_s] = \delta(t-s)$, they are mutually uncorrelated:

$$\begin{aligned}
\mathbf{E}[Z_m Z_n] &= \int_0^1 \int_0^1 \mathbf{E}[\xi_t \xi_s] \psi_m(t) \psi_n(s)\,\mathrm{d}s\,\mathrm{d}t && (2.1.4)\\
&= \int_0^1 \psi_m(t)\psi_n(t)\,\mathrm{d}t && (2.1.5)\\
&= \langle \psi_m, \psi_n \rangle && (2.1.6)\\
&= \delta_{mn} && (2.1.7)
\end{aligned}$$

Finally, since the white noise ξ is a Gaussian process and (2.1.3) expresses each Z_n as a limit of linear operations on ξ, the coefficients Z_n are *Gaussian* random variables. Putting together all of the above, we see that the coefficients $Z_0, Z_1, \ldots$ are independent, identically distributed (i.i.d.) standard Gaussian (zero mean, unit variance) random variables. This suggests that we should be able to obtain a Brownian motion W by passing the expansion (2.1.2) of the white noise ξ through an integrator: $W_t = \int_0^t \xi_s\,\mathrm{d}s$. This is indeed the case:

Theorem 1. *Let $(\psi_n)_{n\in\mathbb{Z}_+}$ be a complete orthonormal basis of $L^2[0,T]$, and let $(Z_n)_{n\in\mathbb{Z}_+}$ be a sequence of i.i.d. standard Gaussian random variables. Then the infinite series*

$$W_t = \sum_{n=0}^{\infty} Z_n \Psi_n(t), \qquad \textit{where } \Psi_n(t) := \int_0^t \psi_n(s)\,\mathrm{d}s \tag{2.1.8}$$

converges uniformly a.s., and the limit is a Brownian motion.

Remark 1. *The phrase "almost sure uniform convergence" here means that, with probability one,*

$$\lim_{N\to\infty} \sup_{t\in[0,1]} |W_t^n - W_t| = 0. \tag{2.1.9}$$

where we have defined the partial sums

$$W_t^n := \sum_{i=0}^{n} Z_i \Psi_i(t). \tag{2.1.10}$$

Note that each W_t^n, being a finite linear combination of i.i.d. standard Gaussian random variables, is itself a Gaussian random variable with zero mean and variance

$$\mathbf{E}[(W_t^n)^2] = \sum_{i=0}^{n} |\Psi_i(t)|^2. \tag{2.1.11}$$

Proof (Sketch). For each $t \in [0,1]$, let I_t denote the indicator function of the interval $[0,t]$, i.e., $I_t(s) = 1_{\{s \le t\}}$. Then $I_t \in L^2[0,1]$ and $\Psi_n(t) = \langle I_t, \psi_n \rangle$. Therefore, for $m, n \in \mathbb{Z}_+$ and $t \in [0,1]$ we have

$$\mathbf{E}[(W_t^m - W_t^n)^2] = \mathbf{E}\left[\left(\sum_{i=n+1}^{m} Z_i \langle I_t, \psi_i \rangle\right)^2\right] \tag{2.1.12}$$

$$= \sum_{i=n+1}^{m} |\langle I_t, \psi_i \rangle|^2 \tag{2.1.13}$$

$$\xrightarrow{m,n\to\infty} 0. \tag{2.1.14}$$

That is, for each t, $(W_t^n)_{n \in \mathbb{Z}_+}$ is a Cauchy sequence in quadratic mean and therefore has a limit W_t, both in quadratic mean and almost surely. Since each W_t^n is Gaussian, so is W_t. It is also straightforward if tedious to show that $(W_t)_{t\in[0,1]}$ is a Gaussian process by working with the limits of the joint characteristic functions of W_t^n's. Finally, using Parseval's relation, we can write

$$\mathbf{E}[W_t W_s] = \mathbf{E}\left[\left(\sum_{m=0}^{\infty} Z_m \Psi_m(t)\right)\left(\sum_{n=0}^{\infty} Z_n \Psi_n(s)\right)\right] \tag{2.1.15}$$

$$= \sum_{n=0}^{\infty} \langle I_t, \psi_n \rangle \langle I_s, \psi_n \rangle \tag{2.1.16}$$

$$= \langle I_t, I_s \rangle \tag{2.1.17}$$

$$= t \wedge s. \tag{2.1.18}$$

Thus, the limit $W = (W_t)_{t\in[0,1]}$ is a Gaussian process that has the properites 1 and 2 of the Brownian motion. It remains to show that it has a.s. continuous sample paths.

While the a.s. uniform convergence holds for any choice of the basis functions ψ_n, the proof in the general case is rather difficult. We will give a direct construction, due to Lévy and Ciesielski, that makes use of a particular basis, the Haar wavelet basis. To introduce this basis, it is convenient to work with two integer indices instead of one; apart from the straightforward relabeling, the ideas are exactly the same. The Haar wavelets $\psi_{k,n}$ with $n = 0, 1, 2, \ldots$ and $k = 1, 3, 5, \ldots, 2^n - 1$ are defined as follows:

$$\psi_{1,0}(t) := 1, \qquad \psi_{k,n}(t) := \begin{cases} +2^{(n-1)/2}, & (k-1)2^{-n} \le t < k2^{-n} \\ -2^{(n-1)/2}, & k2^{-n} \le t < (k+1)2^{-n} \\ 0, & \text{otherwise} \end{cases}. \tag{2.1.19}$$

It is not hard to show that they form a complete orthonormal basis of $L^2[0,1]$. Let us also generate a doubly indexed sequence $(Z_{k,n} : n = 0, 1, \ldots; k = 1, 3, \ldots, 2^n - 1)$ of i.i.d. standard Gaussian random variables. We now define

$$W_t^n := \sum_{i=0}^{n} \sum_{k=1,3,\ldots,2^n-1} Z_{k,i} \Psi_{k,i}(t), \tag{2.1.20}$$

$$W_t := \sum_{n=0}^{\infty} \sum_{k=1,3,\ldots,2^n-1} Z_{k,n} \Psi_{k,n}(t), \tag{2.1.21}$$

where the *Schauder functions* $\Psi_{k,n}(t) = \int_0^t \psi_{k,n}(s)\,\mathrm{d}s$ are given by

$$\Psi_{1,0}(t) = t, \qquad \Psi_{k,n}(t) = \begin{cases} 2^{(n-1)/2}(t-(k-1)2^{-n}), & (k-1)2^{-n} \le t < k2^{-n} \\ (k+1)2^{-(n+1)/2} - 2^{(n-1)/2}t, & k2^{-n} \le t < (k+1)2^{-n} \\ 0, & \text{otherwise} \end{cases}. \tag{2.1.22}$$

So, for $n \ge 1$, $\Psi_{k,n}$ are little 'tents' of height $2^{-(n+1)/2}$ supported on the intervals $[k2^{-n}, (k+1)2^{-n}]$, which are disjoint for distinct values of $k = 1, 3, \ldots, 2^n - 1$. Therefore,

$$\begin{aligned} \sup_{t\in[0,1]} |W_t^n - W_t^{n-1}| &= \sup_{t\in[0,1]} \left| \sum_{k=1,3,\ldots,2^n-1} Z_{k,n}\Psi_{k,n}(t) \right| \\ &= 2^{-(n+1)/2} \max_{k=1,3,\ldots,2^n-1} |Z_{k,n}|. \end{aligned} \tag{2.1.23}$$

Since the maximum of m independent standard Gaussian random variables is on the order of $\sqrt{\log m}$, we see that the maximum of $|W_t^n - W_t^{n-1}|$ over $t \in [0,1]$ is on the order of $\sqrt{n}2^{-n}$, so we expect the series $\sum_{n\ge 0} \sup_{t\in[0,1]} |W_t^n - W_t^{n-1}|$ to be summable; from this, it will be easy to show that $W_t^n \xrightarrow{n\to\infty} W_t$ uniformly in $t \in [0,1]$, and that the limit is a.s. continuous. We now make this argument precise. Using (2.1.23), for any $r > 0$ we can write

$$\begin{aligned} &\mathbf{P}\left[\sup_{t\in[0,1]} |W_t^n - W_t^{n-1}| > r\sqrt{2^{-n}n\log 2} \right] \\ &\quad = \mathbf{P}\left[\max_{k=1,3,\ldots,2^n-1} |Z_{k,n}| > r\sqrt{2n\log 2} \right] && (2.1.24) \\ &\quad \le \sum_{k=1,3,\ldots,2^n-1} \mathbf{P}\left[|Z_{k,n}| > r\sqrt{2n\log 2} \right] && (2.1.25) \\ &\quad \le 2^n \cdot \mathbf{P}[Z > r\sqrt{2n\log 2}], && (2.1.26) \end{aligned}$$

where Z is a standard Gaussian random variable. We now use the Gaussian tail bound $\mathbf{P}[Z > r] \le e^{-r^2/2}$ to write

$$\mathbf{P}\left[\sup_{t\in[0,1]} |W_t^n - W_t^{n-1}| > r\sqrt{2^{-n}n\log 2} \right] \le 2^n \exp\left(-nr^2\log 2\right) \le 2^{n(1-r^2)}, \tag{2.1.27}$$

so, if we take $r > 1$, then

$$\sum_{n=1}^{\infty} \mathbf{P}\left[\sup_{t\in[0,1]} |W_t^n - W_t^{n-1}| > r\sqrt{2^{-n}n\log 2} \right] \le \sum_{n=1}^{\infty} 2^n \exp\left(-nr^2\log 2\right) \le 2^{n(1-r^2)} < \infty, \tag{2.1.28}$$

and therefore

$$\mathbf{P}\left[\sup_{t\in[0,1]} |W_t^n - W_t^{n-1}| \text{ for all sufficiently large } n \right] = 1 \tag{2.1.29}$$

by the Borel–Cantelli lemma. Therefore, almost surely

$$\sum_{n=1}^{\infty} \sup_{t\in[0,1]} |W_t^n - W_t^{n-1}| < \infty. \tag{2.1.30}$$

Since each of the random functions $t \mapsto W_t^n$ is continuous, Lemma 7 in Appendix B tells us that the sequence W^n converges uniformly a.s. to W, which is itself continuous. □

2.1.2 Basic properties

Although Brownian motion has continuous sample paths, they are nowhere differentiable. The proof of this fact is somewhat involved, but the basic intuition is that, since the increments $W_{t+h} - W_t$ for $h > 0$ are zero-mean Gaussian with variance h, they are on the order of $\sqrt{h}$, so the quotient $h^{-1}(W_{t+h} - W_t)$ will blow up as $h \downarrow 0$. In fact, the length (or the total variation) of W is infinite on any interval $[0, t]$. To see this, let us first examine its *quadratic variation* on $[0, t]$. To that end, consider the sequence of random variables

$$Q_t^n := \sum_{0]\le k<2^n t} (W_{(k+1)2^{-n}} - W_{k2^{-n}})^2, \qquad n = 0, 1, \ldots. \tag{2.1.31}$$

By the properties of Brownian motion, Q_t^n is a sum of independent random variables, each with mean $\frac{1}{2^n}$ and variance $\frac{3}{2^{2n}}$. (This follows from the fact that, for $Z \sim N(0, \sigma^2)$, $\mathbf{E}[Z^2] = \sigma^2$ and $\mathbf{E}[Z^4] = 3\sigma^4$.) Therefore, if we consider the sequence $t_n := \max\{k2^{-n} < t\}$, which evidently converges to t, we can write

$$\begin{aligned}
\mathrm{Var}[Q_t^n] &= \mathbf{E}[(Q_t^n - t_n)^2] && (2.1.32)\\
&= 2^n t_n \left(\frac{3}{2^{2n}} - \frac{1}{2^{2n}}\right) && (2.1.33)\\
&= \frac{2t_n}{2^n} && (2.1.34)\\
&< \frac{2t}{2^n}, && (2.1.35)
\end{aligned}$$

so, for any $\varepsilon > 0$, by Chebyshev's inequality

$$\sum_{n=0}^{\infty} \mathbf{P}[|Q_t^n - t_n| \ge \varepsilon] \le \sum_{n=0}^{\infty} \frac{\mathrm{Var}[Q_t^n]}{\varepsilon^2} < \infty. \tag{2.1.36}$$

Thus, by the Borel–Cantelli lemma, $|Q_t^n - t_n| < \varepsilon$ for all but finitely many values of n a.s.; since $\varepsilon > 0$ was arbitrary, it follows that Q_t^n a.s. converges to t as $n \to \infty$. The fact that W has infinite variation on $[0, t]$ can now be derived as follows. First of all, it is easy to see that

$$Q_t^n \le \max_{0\le k<2^n t} |W_{(k+1)2^{-n}} - W_{k2^{-n}}| \cdot \sum_{0\le k<2^n t} |W_{(k+1)2^{-n}} - W_{k2^{-n}}|. \tag{2.1.37}$$

As we have already shown, in the limit as $n \to \infty$, the left-hand side converges to t; moreover, since W has continuous sample paths, $\max_{0\le k<2^n t} |W_{(k+1)2^{-n}} - W_{k2^{-n}}|$ converges a.s. to 0 as $n \to \infty$. Therefore, the sum on the right-hand side has to diverge to infinity as $n \to \infty$.

A key property of Brownian motion is the following. Suppose that we have been observing the evolution of W over the interval $[0, s]$, i.e., we have recorded the trajectory $W_{[0,s]} := (W_r)_{t\in[0,s]}$, and we wish to *predict* the value W_t as some $t > s$. Our prediction, which we denote by $\hat{W}_t$, will be a function of the observations $W_{[0,s]}$, and we want it to be optimal in the mean-square sense, i.e.,

$$\mathbf{E}[(\hat{W}_t - W_t)^2] = \inf_F \mathbf{E}[(F(W_{[0,s]}) - W_t)^2], \tag{2.1.38}$$

where (leaving aside various technicalities) the infimum is over all "reasonable" functions F that map $W_{[0,s]}$ to a prediction of W_t. The minimum mean-square predictor of W_t is given by the conditional expectation of W_t given the observations, i.e.,

$$\hat{W}_t = \mathbf{E}[W_t|W_{[0,s]}], \tag{2.1.39}$$

which we can write, using a convenient shorthand, as $\hat{W}_t = \mathbf{E}[W_t|\mathcal{F}_s]$, where $\mathcal{F}_s$ is the σ-algebra generated by $W_{[0,s]}$. (Roughly speaking, we can think of $\mathcal{F}_s$ as all the "information" about $W_{[0,s]}$.) Then it is not hard to see that $\hat{W}_t = \mathbf{E}[W_t|\mathcal{F}_s] = W_s$:

$$\begin{aligned}\mathbf{E}[W_t|\mathcal{F}_s] &= \mathbf{E}[W_t - W_s + W_s|\mathcal{F}_s] &(2.1.40)\\ &= \mathbf{E}[W_t - W_s|\mathcal{F}_s] + \mathbf{E}[W_s|\mathcal{F}_s] &(2.1.41)\\ &= W_s, &(2.1.42)\end{aligned}$$

where we have used the fact that, for $t > s$, the increment $W_t - W_s$ is independent of $W_{[0,s]}$ and has zero mean, while $\mathbf{E}[W_s|\mathcal{F}_s] = W_s$ because W_s is a deterministic function of the available information $W_{[0,s]}$. This property has a name: We say that the Brownian motion is a *martingale*, a random process with the property that the best prediction (in the minimum mean-square sense) of its value at some time t given the past observations of the process for all times $0 \le r \le s$ is equal to the most recent observation W_s.

Another key property of Brownian motion is that it is a *Markov process*, i.e., for any selection of times $t_1 < t_2 < \cdots < t_{n-1} < t_n$, the conditional probability distribution of W_{t_n} given $(W_{t_i} : 1 \le i < n)$ is equal to the conditional distribution of W_{t_n} given just $W_{t_{n-1}}$. This can be derived from the properties of Brownian motion: By expressing the joint distribution of $(W_{t_i} : 1 \le i \le n)$ in terms of the joint distribution of $(W_{t_1}, W_{t_2} - W_{t_1}, \ldots, W_{t_n} - W_{t_{n-1}})$ and using the fact that the increments $W_{t_i} - W_{t_{i-1}}$ are independent and Gaussian with zero mean and variance $t_i - t_{i-1}$, for any $a < b$ and any $x_1, \ldots, x_{n-1}$, we have

$$\begin{aligned}\mathbf{P}[a \le W_{t_n} \le b|W_{t_i} = x_i, 1 \le i < n] &= \frac{1}{\sqrt{2\pi(t_n - t_{n-1})}} \int_a^b \exp\left(-\frac{(y - x_{n-1})^2}{2(t_n - t_{n-1})}\right) \mathrm{d}y &(2.1.43)\\ &= \mathbf{P}[a \le W_{t_n} \le b|W_{t_{n-1}} = x_{n-1}]. &(2.1.44)\end{aligned}$$

This property can be stated in terms of *transition densities*: For any $x \in \mathbb{R}$, any $0 \le s < t$, and any bounded measurable function $\varphi : \mathbb{R} \to \mathbb{R}$,

$$\begin{aligned}\mathbf{E}[\varphi(W_t)|W_s = x] &= \frac{1}{\sqrt{2\pi(t-s)}} \int_{-\infty}^{\infty} \varphi(y) \exp\left(-\frac{(y-x)^2}{2(t-s)}\right) \mathrm{d}y &(2.1.45)\\ &=: \int_{-\infty}^{\infty} \varphi(y) p_{s,t}(x, y)\, \mathrm{d}y, &(2.1.46)\end{aligned}$$

where $p_{s,t}(x,\cdot)$, for $0 \le s < t$, is the conditional probability density of W_t given $W_s = x$. But W has stationary increments, so it suffices to work with

$$p_t(x,y) := p_{0,t}(x,y) \equiv \frac{1}{\sqrt{2\pi t}} e^{-\frac{(y-x)^2}{2t}} \tag{2.1.47}$$

since evidently $p_{s,t}(x,y) = p_{t-s}(x,y)$. For example, because $W_0 = 0$,

$$\mathbf{E}[\varphi(W_t)] = \int_{-\infty}^{\infty} \varphi(y) p_t(0,y)\,\mathrm{d}y; \tag{2.1.48}$$

more generally,

$$\mathbf{E}[\varphi(W_t)|W_s = x] = \int_{-\infty}^{\infty} \varphi(y) p_{t-s}(x,y)\,\mathrm{d}y. \tag{2.1.49}$$

Let us see what happens when we differentiate both side of this equality w.r.t. the time t:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{E}[\varphi(W_t)|W_s = x] = \int_{-\infty}^{\infty} \varphi(y) \frac{\partial}{\partial t} p_{t-s}(x,y)\,\mathrm{d}y \tag{2.1.50}$$

$$= \int_{-\infty}^{\infty} \varphi(y) \left(\left(\frac{y}{t-s} \right)^2 - \frac{1}{t-s} \right) p_{t-s}(x,y)\,\mathrm{d}y \tag{2.1.51}$$

$$= \frac{1}{2} \int_{-\infty}^{\infty} \varphi(y) \frac{\partial^2}{\partial y^2} p_{t-s}(x,y)\,\mathrm{d}y. \tag{2.1.52}$$

Since φ is arbitrary, we conclude that the transition density $p_{t-s}(x,y)$, as a function of the 'forward space' variable y and the 'forward time' variable t, is a solution of the second-order linear partial differential equation (PDE)

$$\frac{\partial}{\partial t} p_{t-s}(x,y) = \frac{1}{2} \frac{\partial^2}{\partial y^2} p_{t-s}(x,y), \qquad t > s \tag{2.1.53}$$

with the initial condition $\lim_{t \to s} p_{t-s}(x,y) = \delta(y-x)$. Integrating by parts twice and assuming that φ is twice differentiable and sufficiently well-behaved at infinity, we can write

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{E}[\varphi(W_t)|W_s = x] = \int_{-\infty}^{\infty} \varphi''(y) p_{t-s}(x,y)\,\mathrm{d}y \tag{2.1.54}$$

$$= \frac{1}{2}\mathbf{E}[\varphi''(W_t)|W_s = x], \tag{2.1.55}$$

or, in integral form,

$$\mathbf{E}[\varphi(W_t)|W_s = x] = \varphi(x) + \frac{1}{2} \int_s^t \mathbf{E}[\varphi''(W_r)|W_s = x]\,\mathrm{d}r. \tag{2.1.56}$$

2.1.3 Multidimensional Brownian motion

We will be considering processes whose sample paths evolve in the d-dimensional space $\mathbb{R}^d$. The basic object will be the standard d-dimensional Brownian motion, which is simply a vector-valued process $(W_t)_{t \ge 0}$, where $W_t = (W_t^1, \ldots, W_t^d)^T$ is a vector of d independent scalar Brownian motions.

All the properties carry over straightforwardly from the one-dimensional case. Among other things, W has the Gaussian transition densities

$$p_t(x,y) = \frac{1}{(2\pi t)^{d/2}} \exp\left(-\frac{|y-x|^2}{2t}\right), \qquad x,y \in \mathbb{R}^d, t > 0 \tag{2.1.57}$$

where $|x| = \sqrt{x^T x}$ denotes the Euclidean norm of the vector $x = (x^1, \dots, x^d)^T$. These densities are the solutions of the d-dimensional heat equation

$$\frac{\partial}{\partial t} p_t(x,y) = \frac{1}{2} \Delta_y p_t(x,y) \tag{2.1.58}$$

with the initial condition $\lim_{t\to 0} p_t(x,y) = \delta(y-x)$, where $\Delta_y = (\nabla_y \cdot \nabla_y)$ is the Laplace operator and the partial differentiation is w.r.t. the coordinates of y. Consequently, for any twice-differentiable function $\varphi : \mathbb{R}^d \to \mathbb{R}$ well-behaved at infinity,

$$\mathbf{E}[\varphi(W_t)|W_s] = \varphi(W_s) + \frac{1}{2}\int_s^t \mathbf{E}[\Delta\varphi(W_r)|W_s]\,\mathrm{d}r, \qquad t > s. \tag{2.1.59}$$

2.2 Diffusion processes

We can extract the following from the discussion in Section 2.1: If $W = (W_t)_{t\geq 0}$ is the standard d-dimensional Brownian motion, then it has continuous sample paths and, for all $t \geq 0$ and all $h > 0$, we have

$$\mathbf{E}[W_{t+h} - W_t|\mathcal{F}_t] = 0 \tag{2.2.1}$$

$$\mathbf{E}[(W_{t+h} - W_t)(W_{t+h} - W_t)^T|\mathcal{F}_t] = I_d \tag{2.2.2}$$

where I_d is the $d \times d$ identity matrix and $\mathcal{F}_t$ is the σ-algebra generated by the trajectory $(W_r : 0 \leq r \leq t)$ Moreover, using the properties of Gaussian random vectors, it can be shown that

$$\lim_{h\downarrow 0} \frac{1}{h}\mathbf{P}[|W_{t+h} - W_t| > r|\mathcal{F}_t] = 0, \qquad \forall r > 0. \tag{2.2.3}$$

In fact, it can be shown that a process $(W_t)_{t\geq 0}$ with continuous sample paths that has these properties is a Brownian motion.

Generalizing these observations led Andrey Kolmogorov to the following definition: A d-dimensional random process $(X_t)_{t\geq 0}$ with continuous sample paths is a (time-homogeneous) Markov diffusion process if, for any $t \geq 0$ and any $h > 0$,

$$\lim_{h\downarrow 0} \frac{1}{h}\mathbf{P}[|X_{t+h} - X_t| > r|\mathcal{F}_t] = 0, \qquad \forall r > 0 \tag{2.2.4}$$

and if there exist a vector-valued function $f : \mathbb{R}^d \to \mathbb{R}^d$ (the *drift*) and a matrix-valued function $a : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ (the *diffusion matrix*), such that, for all $t \geq 0$ and all $h > 0$,

$$\begin{aligned} \lim_{h\downarrow 0} \frac{1}{h}\mathbf{E}[X_{t+h} - X_t|\mathcal{F}_t] &= f(X_t) \\ \lim_{h\downarrow 0} \frac{1}{h}\mathbf{E}[(X_{t+h} - X_t)(X_{t+h} - X_t)^T|\mathcal{F}_t] &= a(X_t). \end{aligned} \tag{2.2.5}$$

It follows from the above that $a(x)$ is symmetric and positive semidefinite for each $x \in \mathbb{R}^d$. The Brownian motion is a special case of this, with $f(x) \equiv 0$ and $a(x) \equiv I_d$.

Theorem 2. *Let $(X_t)_{t\geq 0}$ be a d-dimensional Markov diffusin process with drift f and diffusion matrix a. Then, for any twice differentiable function $\varphi : \mathbb{R}^d \to \mathbb{R}$,*

$$\mathbf{E}[\varphi(X_t)|X_s = x] = \varphi(x) + \int_s^t \mathbf{E}[\mathcal{A}\varphi(X_r)|X_s = x]\,\mathrm{d}r, \tag{2.2.6}$$

where $\mathcal{A}$ is the linear second-order differential operator given by

$$\mathcal{A}\varphi(x) := f(x)^T\nabla\varphi(x) + \frac{1}{2}\mathrm{tr}\{a(x)\nabla^2\varphi(x)\}. \tag{2.2.7}$$

The operator $\mathcal{A}$ is called the *infinitesimal generator* of the diffusion process. While the full rigorous proof is beyond the scope of these notes, we can give a sketch of the main idea. Define the function $F_s(x,t) := \mathbf{E}[\varphi(X_t)|X_s = x]$ for $t \geq s$ and $x \in \mathbb{R}^d$, and fix some $h > 0$. Then, expanding $\varphi(X_{t+h})$ in a Taylor series around X_t, we can write

$$\begin{aligned}
&F_s(x,t+h) = \mathbf{E}[\varphi(X_{t+h})|X_s = x] \\
&= \mathbf{E}\Big[\varphi(X_t) + \nabla\varphi(X_t)^T(X_{t+h} - X_t) + \frac{1}{2}(X_{t+h} - X_t)^T\nabla^2\varphi(X_t)(X_{t+h} - X_t) + R_{t,h}\Big|X_s = x\Big] &(2.2.8)\\
&= F_s(x,t) + \mathbf{E}\Big[\nabla\varphi(X_t)^T(X_{t+h} - X_t) + \frac{1}{2}(X_{t+h} - X_t)^T\nabla^2\varphi(X_t)(X_{t+h} - X_t) + R_{t,h}\Big|X_s = x\Big] &(2.2.9)
\end{aligned}$$

where $R_{t,h}$ is the remainder term that scales as $o(|X_{t+h} - X_t|^2)$. Using the law of iterated expectation, the fact that $\mathcal{F}_t \supseteq \mathcal{F}_s$, and (2.2.5), we can write

$$\begin{aligned}
&\mathbf{E}\Big[\nabla\varphi(X_t)^T(X_{t+h} - X_t) + \frac{1}{2}(X_{t+h} - X_t)^T\nabla^2\varphi(X_t)(X_{t+h} - X_t) + R_{t,h}\Big|X_s = x\Big] &(2.2.10)\\
&= \mathbf{E}\Big[\mathbf{E}\Big[\nabla\varphi(X_t)^T(X_{t+h} - X_t) + \frac{1}{2}(X_{t+h} - X_t)^T\nabla^2\varphi(X_t)(X_{t+h} - X_t) + R_{t,h}\Big|\mathcal{F}_t\Big]\Big|X_s = x\Big] &(2.2.11)\\
&= \mathbf{E}\Big[\nabla\varphi(X_t)^T(hf(X_t) + o(h)) + \frac{1}{2}\mathrm{tr}\{\nabla^2\varphi(X_t)(ha(X_t) + o(h)) + R_{t,h}\Big|X_s = x\Big]. &(2.2.12)
\end{aligned}$$

Rearranging and using the definition of $\mathcal{A}$ gives

$$\frac{1}{h}(F_s(x,t+h) - F_s(x,t)) = \mathbf{E}[\mathcal{A}\varphi(X_t)|X_s = x] + \frac{1}{h}\mathbf{E}[R_{t,h}|X_s = x] + \frac{o(h)}{h}. \tag{2.2.13}$$

Taking the limit as $h \downarrow 0$, we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{E}[\varphi(X_t)|X_s = x] = \mathbf{E}[\mathcal{A}\varphi(X_t)|X_s = x] + \lim_{h\downarrow 0}\frac{1}{h}\mathbf{E}[R_{t,h}|X_s = x]. \tag{2.2.14}$$

Using (2.2.4) it can be shown that the last term on the right-hand side vanishes, and we get (2.2.6) by integrating.

2.3 The Kolmogorov equations

So far, Brownian motion is the only diffusion process whose existence we have proved constructively. While its definition is framed in the externalist picture, the explicit construction using wavelets provides the complementary internalist description. This is not the case with general diffusion processes: Kolmogorov's definition is purely externalist, dealing as it does with conditional expectations, and also local in time and in space, prescribing the first- and second-order statistics of the increments of the process on infinitesimally small time intervals and in arbitrarily small neighborhoods of the current state of the process. In order to give the complementary internalist description of a general diffusion process, we will need the machinery of Itô calculus, and it will turn out that all such processes can be built up from Brownian motion. For now, though, we will show that the linear second-order differential operator $\mathcal{A}$ in (2.2.7) plays a key role in constraining the behavior of the transition probabilities of the process.

As a reminder, any time-homogeneous continuous-time Markov process $(X_t)_{t\geq 0}$ admits a probabilistic description in terms of its transition probability functions $P_t(x, A)$, for $t > 0$, $x \in \mathbb{R}^d$, and Borel sets A:

$$\mathbf{P}[X_{s+t} \in A | X_s = x] = P_t(x, A) \tag{2.3.1}$$

or, more generally,

$$\mathbf{E}[\varphi(X_{s+t})|X_s = x] = \int_{\mathbb{R}^d} \varphi(y) P_t(x, \mathrm{d}y). \tag{2.3.2}$$

The transition probability functions obey the Champan–Kolmogorov equation,

$$P_t(x, A) = \int_{\mathbb{R}^d} P_{t-s}(y, A) P_s(x, \mathrm{d}y), \qquad 0 < s < t \tag{2.3.3}$$

which expresses mathematically the intuitive picture that, in order to compute the probability that the process reaches the set A from the point x in time t, we need to consider the probabilities of intermediate positions after some fixed time $s < t$. Under appropriate regularity conditions, the transition probability functions can be expressed in terms of transition densities $p_t(x, y)$:

$$P_t(x, A) = \int_A p_t(x, y)\, \mathrm{d}y \tag{2.3.4}$$

where $p_t(x, \cdot)$ is the conditional probability density of X_{s+t} given $X_s = x$ for any $s \geq 0$. While establishing the existence of transition densities is a delicate matter, we will sweep these technical issues under the rug and, assuming that the diffusion process of interest is sufficiently well-behaved to admit transition densities, derive two PDEs that they must satisfy. These are known as the forward and the backward Kolmogorov equations.

We will assume that the transition densities $p_t(x, y)$ are once continuously differentiable in t and twice continuously differentiable in x and y, and that the first and second derivatives of $p_t(x, y)$ w.r.t. x are continuous in t. We will start by deriving the forward Kolmogorov equation. Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be an infinitely differentiable function supported on a compact subset of $\mathbb{R}^d$. Then, on the one hand we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{E}[\varphi(X_t)|X_s = x] = \int_{\mathbb{R}^d} \varphi(y) \frac{\partial}{\partial t} p_{t-s}(x, y)\, \mathrm{d}y. \tag{2.3.5}$$

On the other hand, Theorem 2 tells us that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{E}[\varphi(X_t)|X_s = x] = \mathbf{E}[\mathcal{A}\varphi(X_t)|X_s = x]. \tag{2.3.6}$$

Using the definition (2.2.7) of $\mathcal{A}$ and integrating by parts, we get

$$\begin{aligned}
&\mathbf{E}[\mathcal{A}\varphi(X_t)|X_s = x] \\
&= \int_{\mathbb{R}^d} \mathcal{A}\varphi(y) p_{t-s}(x,y)\,\mathrm{d}y && (2.3.7)\\
&= \int_{\mathbb{R}^d} f(y)^T \nabla\varphi(y) p_{t-s}(x,y)\,\mathrm{d}y + \frac{1}{2}\int_{\mathbb{R}^d} \mathrm{tr}\{a(y)\nabla^2\varphi(y)\} p_{t-s}(y)\,\mathrm{d}y && (2.3.8)\\
&= \int_{\mathbb{R}^d} \left\{ \sum_{i=1}^d \left(\frac{\partial}{\partial y^i}\varphi(y)\right) f_i(y) + \frac{1}{2}\sum_{i,j=1}^d \left(\frac{\partial^2}{\partial y^i \partial y^j}\varphi(y)\right) a_{ij}(y) \right\} p_{t-s}(x,y)\,\mathrm{d}y && (2.3.9)\\
&= \int_{\mathbb{R}^d} \varphi(y) \left\{ -\sum_{i=1}^d \frac{\partial}{\partial y^i}\left(f_i(y) p_{t-s}(x,y)\right) + \frac{1}{2}\sum_{i,j=1}^d \frac{\partial^2}{\partial y^i \partial y^j}\left(a_{ij}(y) p_{t-s}(x,y)\right) \right\} \mathrm{d}y. && (2.3.10)
\end{aligned}$$

The right-hand sides of (2.3.5) and (2.3.10) are equal, and, since φ was arbitrary, it must be the case that the transition density $p_{t-s}(x,y)$ is a solution of the PDE

$$\frac{\partial}{\partial t} p_{t-s}(x,y) = -\sum_{i=1}^d \frac{\partial}{\partial y^i}\left(f_i(y) p_{t-s}(x,y)\right) + \frac{1}{2}\sum_{i,j=1}^d \frac{\partial^2}{\partial y^i \partial y^j}\left(a_{ij}(y) p_{t-s}(x,y)\right), \qquad t > s \tag{2.3.11}$$

with the initial condition $\lim_{t\to s} p_{t-s}(x,y) = \delta(y-x)$. This is *Kolmogorov's forward equation* (also known as the Fokker–Planck equation); we can also use it to describe the evolution of the marginal density ρ_t of X_t when the initial state X_0 has a given density ρ_0. In terms of the transition densities, we have

$$\rho_t(y) = \int_{\mathbb{R}^d} \rho_0(x) p_t(x,y)\,\mathrm{d}x; \tag{2.3.12}$$

differentiating both sides w.r.t. t and using (2.3.11), we get

$$\frac{\partial}{\partial t}\rho_t(y) = -\sum_{i=1}^d \frac{\partial}{\partial y^i}\left(f_i(y)\rho_t(y)\right) + \frac{1}{2}\sum_{i,j=1}^d \frac{\partial^2}{\partial y^i \partial y^j}\left(a_{ij}(y)\rho_t(y)\right), \qquad t > 0 \tag{2.3.13}$$

with the given initial condition ρ_0 at $t = 0$.

To derive Kolmogorov's backward equation, we start with the Chapman–Kolmogorov equation for the transition densities: for $0 \le s < t$ and $x, y \in \mathbb{R}^d$,

$$p_{t-s}(x,y) = \int_{\mathbb{R}^d} p_{t-s-h}(z,y) p_h(x,z)\,\mathrm{d}z, \tag{2.3.14}$$

where $h \in \mathbb{R}$ is sufficiently small, so that $t - s > h$. Expanding the function $z \mapsto p_{t-s-h}(z,y)$ to second order in a neighborhood of x gives

$$\begin{aligned}
p_{t-s-h}(z,y) = p_{t-s-h}(x,y) &+ \nabla_x p_{t-s-h}(x,y)^T (z-x) \\
&+ \frac{1}{2}(z-x)^T \nabla_x^2 p_{t-s-h}(x,y)(z-x) + o(|x-z|^2)
\end{aligned} \tag{2.3.15}$$

Substituting this into (2.3.14), we obtain

$$\begin{aligned} p_{t-s}(x,y) &= \int_{\mathbb{R}^d} p_{t-s-h}(x,y)p_h(x,z)\,\mathrm{d}z \\ &\quad + \int_{\mathbb{R}^d} \nabla_x p_{t-s-h}(x,y)^T(z-x)p_h(x,z)\,\mathrm{d}z \\ &\quad + \frac{1}{2}\int_{\mathbb{R}^d} (z-x)^T\nabla_x^2 p_{t-s-h}(x,y)(z-x)p_h(x,z)\,\mathrm{d}z \\ &\quad + o(h) \qquad (2.3.16) \\ &= p_{t-s-h}(x,y) + \nabla_x p_{t-s-h}(x,y)^T\mathbf{E}[X_h - X_0|X_0 = x] \\ &\quad + \frac{1}{2}\mathrm{tr}\left\{\mathbf{E}[(X_h-X_0)(X_h-X_0)^T|X_0=x]\nabla_x^2 p_{t-s-h}(x,y)\right\} + o(h). \qquad (2.3.17) \end{aligned}$$

Rearranging, dividing by h, and using (2.2.5) gives

$$\begin{aligned} &-\frac{1}{h}[p_{t-(s+h)}(x,y) - p_{t-s}(x,y)] \\ &\quad = \frac{1}{h}\nabla_x p_{t-(s+h)}(x,y)^T\mathbf{E}[X_h - X_0|X_0 = x] \\ &\qquad + \frac{1}{2h}\mathrm{tr}\left\{\mathbf{E}[(X_h-X_0)(X_h-X_0)^T|X_0=x]\nabla_x^2 p_{t-(s+h)}(x,y)\right\} + \frac{1}{h}o(h) \qquad (2.3.18) \\ &\quad = \frac{1}{h}\cdot(hf(x)+o(h))^T\nabla_x p_{t-(s+h)}(x,y) \\ &\qquad + \frac{1}{2}\mathrm{tr}\left\{\frac{1}{h}\cdot(ha(x)+o(h))^T\nabla_x^2 p_{t-(s+h)}(x,y)\right\} + \frac{1}{h}o(h). \qquad (2.3.19) \end{aligned}$$

Taking the limit as $h \to 0$ and using the assumed continuity of $t \mapsto p_t(x,y)$, we get the backward Kolmogorov equation

$$\frac{\partial}{\partial s}p_{t-s}(x,y) = -\sum_{i=1}^d f_i(x)\frac{\partial}{\partial x^i}p_{t-s}(x,y) - \frac{1}{2}\sum_{i,j=1}^d a_{ij}(x)\frac{\partial^2}{\partial x^i\partial x^j}p_{t-s}(x,y) \qquad (2.3.20)$$

for $0 \le s \le t$ with the terminal condition $\lim_{s\to t} p_{t-s}(x,y) = \delta(y-x)$. Note that we can rewrite this more succinctly using the definition of $\mathcal{A}$:

$$\frac{\partial}{\partial s}p_{t-s}(x,y) = -\mathcal{A}p_{t-s}(x,y), \qquad (2.3.21)$$

where the derivatives in $\mathcal{A}$ act on x while keeping y fixed.

What we have done is start with a diffusion process characterized by the drift f and the diffusion matrix a and, assuming it is regular enough to admit transition densities, derived two PDEs that must be obeyed by these transition densities. It turns out that we can also go the other way around — that is, if $p_t(x,y)$ are smooth solutions of (2.3.11) and (2.3.20) for a given pair of f and a, then they are the transition densities of a diffusion process with drift f and diffusion matrix a.

2.4 Problems

1. Prove that a scalar zero-mean Gaussian process $(V_t)_{t\ge 0}$ with $V_0 = 0$ has the covariance function $\mathbf{E}[V_sV_t] = s \wedge t$ if and only if the increments $V_t - V_s$ are zero-mean Gaussian with variance $|t-s|$. *Hint:* Use the identity $a \wedge b = \frac{1}{2}(a+b-|a-b|)$.

2. Let $(V_t)_{t\geq 0}$ be a scalar zero-mean Gaussian process with covariance function $\mathbf{E}[V_s V_t] = s \wedge t$. Prove that it has independent increments, i.e., for any n and $0 \leq t_1 < t_2 < \cdots < t_n$, the random variables $V_{t_2} - V_{t_1}, V_{t_3} - V_{t_2}, \ldots, V_{t_n} - V_{t_{n-1}}$ are independent.

3. Let $(W_t)_{t\geq 0}$ be a standard one-dimensional Brownian motion. Prove that the following processes are also Brownian motion processes:

(i) $(W_{t+a} - W_a)_{t\geq 0}$, where $a \geq 0$ is a fixed constant.

(ii) $(\lambda^{-1} W_{\lambda^2 t})_{t\geq 0}$, where $\lambda \neq 0$ is a fixed constant.

(iii) $(tW_{1/t})_{t>0}$.

4. Let $(X_1, \ldots, X_n)$ be an n-dimensional random vector with a joint pdf p_X. Express p_X in terms of the joint pdf of $(X_1, X_2 - X_1, \ldots, X_n - X_{n-1})$. Use this fact to give a proof that Brownian motion is a Markov process with transition densities

$$p_t(x, y) = \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{(y-x)^2}{2t} \right).$$

5. Let $(X_t)_{t\geq 0}$ be a d-dimensional diffusion process. Show that, for $t \geq 0$ and $h > 0$,

$$\mathrm{Cov}[X_{t+h} - X_t | X_t] = ha(X_t) + o(h)$$

6. Let $(W_t)_{t\geq 0}$ be a standard one-dimensional Brownian motion, and let $\varphi : \mathbb{R} \to \mathbb{R}$ be a strictly monotone, twice continuously differentiable function. Show that $X_t = \varphi(W_t)$ is a one-dimensional diffusion process with drift $f(x) = \frac{1}{2}\varphi''(\varphi^{-1}(x))$ and diffusion coefficient $a(x) = |\varphi'(\varphi^{-1}(x))|^2$, where φ^{-1} is the inverse of φ, i.e., the unique solution y to the equation $\varphi(y) = x$.

7. Let $(X_t)_{t\geq 0}$ be a d-dimensional diffusion process, and let $\varphi : \mathbb{R}^d \times [0, \infty) \to \mathbb{R}$ be a function of $x \in \mathbb{R}^d$ and $t \in [0, \infty)$, which is twice continuously differentiable in x and once continuously differentiable in t. Prove that, for any $t > s \geq 0$,

$$\mathbf{E}[\varphi(X_t, t) | \mathcal{F}_s] = \varphi(X_s, s) + \int_s^t \mathbf{E}\left[\dot{\varphi}(X_r, r) + \mathcal{A}\varphi(X_r, r) \middle| \mathcal{F}_s \right] \mathrm{d}r$$

where $\mathcal{F}_s$ is the σ-algebra generated by $(X_r)_{0\leq r\leq s}$, $\dot{\varphi}(x,t) := \frac{\partial}{\partial t}\varphi(x,t)$, and $\mathcal{A}$ is the infinitesimal generator of $(X_t)_{t\geq 0}$, which acts on functions of both x and t as

$$\mathcal{A}\varphi(x,t) = f(x)^T \nabla_x \varphi(x,t) + \frac{1}{2}\mathrm{tr}\{a(x)\nabla_x^2 \varphi(x,t)\}.$$

8. Consider a scalar diffusion process with drift $f(x) = -x$, and diffusion coefficient $a(x) = 2$ (this is known as the *standard Ornstein–Uhlenbeck process* in one dimension). Suppose that X_0 has the standard Gaussian density $\rho_0(x) = (2\pi)^{-1/2} e^{-x^2/2}$. Using the Fokker–Planck equation, show that the marginal densities $\rho_t(x)$ of X_t are all equal to ρ_0.

2.5 Notes and further reading

The construction of the Brownian motion in Section 2.1 largely follows [Cla73].

Chapter 3

Stochastic Integrals and Stochastic Differential Equations

In Chapter 2, we have introduced diffusion processes from an externalist point of view in terms of their drift and diffusion coefficients. This is a local description of a diffusion process which, in fact, completely determines the transition probability functions of the process. These, in turn, allow us to compute the probabilities of various events pertaining to the process, but tell us nothing about the way one could go about generating a process with these transition probabilities. On the other hand, an important first step in building models of physical and engineering systems with random effects is to go from an externalist description to an internalist one involving latent variables not amenable to direct observation. This passage from a given externalist description to a suitable internalist model is often referred to as the *realization problem*, and one typically imposes some structural constraints on the mechanism that relates the internal variables to the external variables.

You have no doubt encountered this type of modeling in the first course on probability theory: Any real-valued random variable X with cdf $F_X(x) = \mathbf{P}[X \leq x]$ can be realized as a function of a random variable U uniformly distributed on the unit interval $[0,1]$, i.e., $F_X^{-1}(U)$ and X have the same cdf. This is readily verified using the fact that the cdf of U is $F_U(u) = u$ for $u \in [0,1]$. In this case, U is the internal or latent variable and X is the external or observed variable. Of course, not every instance of X is necessarily generated in this way, but a representation of this kind is nevertheless very useful. For instance, if we know the cdf F_X, then we can use it as many independent copies of X as we want; if we don't know F_X, then we can attempt to learn it by comparing the cdf of $\hat{F}^{-1}(X)$ with the cdf of U for various candidates $\hat{F}$ — this is precisely the idea behind the Kolmogorov–Smirnov goodness-of-fit test.

Our goal in this chapter is twofold: We will first show that a wide class of diffusion processes can be realized as deterministic functions of Brownian motion. That is, if $(X_t)_{t\geq 0}$ is a d-dimensional diffusion process satisfying certain regularity conditions, then it can be realized as a function of a d-dimensional Brownian motion $(W_t)_{t\geq 0}$, i.e., for each t there exists a mapping F_t such that the entire path $(X_s)_{0\leq s\leq t}$ can be obtained as $F_t[X_0, (W_s)_{0\leq s\leq t}]$, where the (possibly random) starting point X_0 is independent of the Brownian motion. This Brownian motion plays the role of the latent object; moreover, we will be able to write down the form of F_t explicitly in terms of the drift and the diffusion coefficients of $(X_t)_{t\geq 0}$. To that end, we will need to introduce stochastic integrals in the sense of Itô, and this will in turn lead us to the construction of stochastic differential equations. Martingale theory will play an important role here as well. One can think of this result as solving

the realization problem for diffusion processes. We will then turn this idea around — we will *start* with internalist models based on stochastic differential equations and show that, under appropriate regularity conditions, their solutions are in fact diffusion processes. Given this, it is tempting to view stochastic differential equations as limits of ordinary differential equations driven by wideband noise. As we shall see, however, this will run into certain complications having to do with the fact that the rules of Itô calculus are different from the rules of ordinary calculus.

3.1 From Kolmogorov to Itô: an internalist model of a diffusion process

As mentioned above, we will introduce Itô's stochastic integrals in the context of solving the realization problem for diffusion processes. To keep things simple, we will consider the one-dimensional case; the extension to vector-valued processes is relatively straightforward once the main ideas are understood. Suppose that we are observing some stochastic system over a time interval of finite length T. Our measurement apparatus connected to the system produces a random trajectory $\boldsymbol{X} = (X_t)_{t\in[0,T]}$, and we will assume the act of measurement does not introduce any additional disturbances. We also assume that the observed trajectory is a diffusion process with known drift f and diffusion coefficient a, where both are sufficiently regular to ensure that the processes $f(X_t)$ and $a(X_t)$ have continuous sample paths. For instance, since X_t, being a diffusion process, has continuous sample paths and if both f and a are continuous functions of their argument, then the above requirement will be met.

Our goal is to show that $\boldsymbol{X}$ can be represented as a deterministic function of the initial condition X_0 and a Brownian motion $\boldsymbol{W} = (W_t)_{t\in[0,T]}$ independent of X_0 subject to the reasonable causality requirement that, for each $t \in [0,T]$, $(X_\tau)_{\tau\in[0,t]}$ is determined only by X_0 and $(W_\tau)_{\tau\in[0,t]}$.

3.1.1 Filtrations, martingales, and all that

The observed process $\boldsymbol{X}$ is defined on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. The above causality requirement also forces us to consider, for each time $t \in [0,T]$, a smaller σ-algebra $\mathcal{F}_t$ which is a subset of $\mathcal{F}$ and which represents all the information available at time t. Since we are restricted to external observations only, we will take $\mathcal{F}_t$ to be the σ-algebra generated by $(X_\tau)_{\tau\in[0,t]}$. In particular, $\mathcal{F}_t$ contains all events of the form

$$E = \{\omega \in \Omega : a_1 \le X_{t_1}(\omega) < b_1, \ldots, a_n \le X_{t_n}(\omega) < b_n\} \tag{3.1.1}$$

for all finite collections of times $t_1, \ldots, t_n \in [0,t]$ and intervals $[a_1, b_1), \ldots, [a_n, b_n)$, as well as all other events that could be generated from these by taking complements and countable unions. For example, if the system under observation is an electrical circuit that contains noise sources and X_t is the random voltage across a designated pair of terminals at time t, then the event E in (3.1.1) represents all the circumstances under which the voltages measured by an ideal voltmeter at times $0 \le t_1, \ldots, t_n \le t$ fall within given ranges $[a_1, b_1), \ldots, [a_n, b_n)$. The σ-algebra $\mathcal{F}_0$ contains all available information about the initial condition X_0. It is not hard to see that these σ-algebras are nested: If $0 \le t < t'$, then $\mathcal{F}_0 \subseteq \mathcal{F}_t \subseteq \mathcal{F}_{t'} \subseteq \mathcal{F}$. This represents the situation that we accumulate information as time goes on. Any such increasing collection of σ-algebras is called a *filtration*, and we say that a stochastic process $\boldsymbol{Y} = (Y_t)_{t\in[0,T]}$ is *adapted* to the filtration $(\mathcal{F}_t)_{t\in[0,T]}$ if each Y_t is *measurable* w.r.t. $\mathcal{F}_t$, i.e., the information available at time t completely determines Y_t. In our current setting, each $\mathcal{F}_t$ is generated by $(X_\tau)_{\tau\in[0,t]}$, the process $\boldsymbol{X}$ is evidently adapted to the resulting

filtration (we then say that this filtration is *generated* by $\boldsymbol{X}$). Moreover, we will see that the 'latent' Brownian motion $\boldsymbol{W}$ will also be adapted to the same filtration, which makes intuitive sense in light of our causality requirement. This simply means that, for any $0 \leq s < t \leq T$, the increment $W_t - W_s$ is a zero-mean Gaussian random variable with variance $t - s$, which is also independent of $\mathcal{F}_s$ (in our case, this means that $W_t - W_s$ is independent of $(X_r)_{0 \leq r \leq s}$). In particular, we will have

$$\mathbf{E}[W_t - W_s | \mathcal{F}_s] = 0 \text{ and } \mathbf{E}[(W_t - W_s)^2 | \mathcal{F}_s] = t - s, \qquad t \geq s. \tag{3.1.2}$$

This naturally leads us to another important definition: A stochastic process $(Y_t)_{t \in [0,T]}$ adapted to a filtration $(\mathcal{F}_t)_{t \in [0,T]}$ is a *martingale* w.r.t. this filtration if

$$\mathbf{E}[Y_t - Y_s | \mathcal{F}_s] = 0, \qquad t \geq s. \tag{3.1.3}$$

We will often abbreviate this terminology by saying that $\boldsymbol{Y}$ is an $\mathcal{F}_t$-martingale or that $(Y_t, \mathcal{F}_t)_{t \in [0,T]}$ is a martingale. It is important to keep in mind that a given process may be adapted to many different filtrations, and that it may be a martingale w.r.t. one such filtration, but fail to be a martingale w.r.t. another.

Now, the first equality in (3.1.2) tells us that any Brownian motion adapted to $\mathcal{F}_t$ is an $\mathcal{F}_t$-martingale. An important result, which we will use soon, is that any martingale with continous sample paths that satisfies (3.1.2) is, in fact, a Brownian motion:

Theorem 3 (Lévy's characterization of Brownian motion)**.** *Let $\boldsymbol{W} = (W_t)_{t \geq 0}$ be a martingale w.r.t. $(\mathcal{F}_t)_{t \geq 0}$ with continuous sample paths, such that $W_0 = 0$ and*

$$\mathbf{E}[(W_t - W_s)^2 | \mathcal{F}_s] = t - s, \qquad t \geq s. \tag{3.1.4}$$

Then ($\boldsymbol{W}$ is an $\mathcal{F}_t$-Brownian motion.

See Example 4 for the proof.

It will be convenient to adopt the notation $\mathrm{d}Y_t$ for the forward differentials $Y_{t+\mathrm{d}t} - Y_t$ for $0 \leq t < T$ and a small $\mathrm{d}t > 0$ such that $t + \mathrm{d}t \leq T$. Then $\boldsymbol{Y}$ is a martingale w.r.t. $\mathcal{F}_t$ if it is adapted to $\mathcal{F}_t$ and if $\mathbf{E}[\mathrm{d}Y_t | \mathcal{F}_t] = 0$ for all $0 \leq t < T$. From now on, we will be working with such forward differentials whenever possible; among other things, this will allow us to neglect terms of order $(\mathrm{d}t)^2$ and higher. In this notation, we can restate the condition (3.1.4) as $\mathbf{E}[(\mathrm{d}W_t)^2 | \mathcal{F}_t] = \mathrm{d}t$.

There are multiple ways to construct martingales adapted to a given filtration $\mathcal{F}_t$. For example, let a random variable V be given. Then the process $Y_t := \mathbf{E}[V | \mathcal{F}_t]$ is an $\mathcal{F}_t$-martingale, an immediate consequence of the properties of conditional expectation. Another way is to start with a process $\boldsymbol{Z} = (Z_t)_{t \in [0,T]}$ which is not necessarily adapted to $\mathcal{F}_t$ and then construct another process $\boldsymbol{A}$ by setting $A_0 = 0$ and

$$\mathrm{d}A_t := \mathbf{E}[\mathrm{d}Z_t | \mathcal{F}_t]. \tag{3.1.5}$$

Then $M_t := Z_t - A_t$ is an $\mathcal{F}_t$-martingale: It is obviously adapted to $\mathcal{F}_t$, and

$$\mathbf{E}[\mathrm{d}M_t | \mathcal{F}_t] = \mathbf{E}[\mathrm{d}Z_t | \mathcal{F}_t] - \mathbf{E}[\mathrm{d}A_t | \mathcal{F}_t] = 0. \tag{3.1.6}$$

For example, if $\mathcal{F}_t$ is the filtration generated by Z_t, i.e., $\mathcal{F}_t = \sigma(Z_\tau : 0 \leq \tau \leq t)$, then $\mathrm{d}M_t = \mathrm{d}Z_t - \mathbf{E}[\mathrm{d}Z_t | Z_\tau, 0 \leq \tau \leq t]$ contains the new information in $\mathrm{d}Z_t$ on top of $\mathcal{F}_t$. The process M_t is then called the *innovations process* for Z_t.

Finally, an important concept is that of *quadratic variation*. We have already seen it in the case of Brownian motion: If $(W_t)_{t\geq 0}$ is a standard Brownian motion, then the limit

$$\lim_{n\to\infty} \sum_{0\leq k<2^n t} \left(W_{(k+1)2^{-n}} - W_{k2^{-n}}\right)^2 = t \tag{3.1.7}$$

exists almost surely; in fact, it does not even depend on the specific sequence of partitions of the interval $[0,t]$, as long as they become increasingly fine in the limit. We can now consider any square-integrable process Y_t (this means that $\mathbf{E}[Y_t^2] < \infty$ for all $t \in [0,T]$) with continuous sample paths and define its quadratic variation on $[0,t]$ as

$$[Y]_t := \lim_{n\to\infty} \sum_{0\leq k<2^n t} \left(Y_{(k+1)2^{-n}} - Y_{k2^{-n}}\right)^2 \tag{3.1.8}$$

(again, we can take the limit along any sequence of partitions of $[0,t]$, as long as they become increasingly fine in the limit). For example, if W_t a standard Brownian motion, then $[W]_t = t$ almost surely. Also, the same argument that was used in Section 2.1 to show that Brownian motion has infinite total variation on any finite interval can be applied to show that any process Y_t whose quadratic variation $[Y]_t$ is a.s. finite and increasing with t (which will be the case whenever Y_t is not a.s. constant) must have infinite total variation on $[0,t]$.

Now, quadratic variation has special properties when the underlying process is a martingale. Thus, let M_t be a square-integrable $\mathcal{F}_t$-martingale with continuous sample paths, and let us consider the process $Y_t := M_t^2$. Then, using the properties of conditional expectations and the fact that M_t is a martingale, we have

$$\begin{aligned}
\mathbf{E}[\mathrm{d}Y_t|\mathcal{F}_t] &= \mathbf{E}[M_{t+\mathrm{d}t}^2 - M_t^2|\mathcal{F}_t] && (3.1.9)\\
&= \mathbf{E}[(M_{t+\mathrm{d}t} - M_t)^2 + 2(M_{t+\mathrm{d}t} - M_t)M_t|\mathcal{F}_t] && (3.1.10)\\
&= \mathbf{E}[(\mathrm{d}M_t)^2|\mathcal{F}_t] + 2\,\mathbf{E}[M_t\,\mathrm{d}M_t|\mathcal{F}_t] && (3.1.11)\\
&= \mathbf{E}[(\mathrm{d}M_t)^2|\mathcal{F}_t] + 2\,M_t\mathbf{E}[\mathrm{d}M_t|\mathcal{F}_t] && (3.1.12)\\
&= \mathbf{E}[(\mathrm{d}M_t)^2|\mathcal{F}_t]. && (3.1.13)
\end{aligned}$$

Then the process $\boldsymbol{A}$ given by $A_0 = 0$ and $\mathrm{d}A_t := \mathbf{E}[\mathrm{d}Y_t|\mathcal{F}_t]$ is nondecreasing since $(\mathrm{d}M_t)^2 \geq 0$, and is in fact precisely the quadratic variation $[M]_t$. In fact, it will be *strictly increasing* unless M_t is a.s. constant. (Note, by the way, that the definition of quadratic variation does not depend on any particular filtration, unlike our definition of A_t.)

3.1.2 A martingale associated with a diffusion process

We will now see why we needed the above interlude on martingales. Getting back to our diffusion process $\boldsymbol{X}$, we can write

$$\mathbf{E}[\mathrm{d}X_t|\mathcal{F}_t] = f(X_t)\,\mathrm{d}t, \qquad 0 \leq t < T \tag{3.1.14}$$

— indeed, this is just the restatement of $\mathbf{E}[X_{t+\mathrm{d}t} - X_t|\mathcal{F}_t] = f(X_t)\,\mathrm{d}t + o(\mathrm{d}t)$ that we have been using before. Then the process $\boldsymbol{M} = (M_t)_{t\in[0,T]}$ defined by $M_0 = 0$ and

$$\mathrm{d}M_t := \mathrm{d}X_t - f(X_t)\,\mathrm{d}t, \tag{3.1.15}$$

or, in integral form,

$$M_t = X_t - X_0 - \int_0^t f(X_s)\,\mathrm{d}s, \tag{3.1.16}$$

is an $\mathcal{F}_t$-martingale. Indeed, as can be easily seen from (3.1.16), M_t is completely determined by $(X_s)_{s\in[0,t]}$, i.e., it is measurable w.r.t. $\mathcal{F}_t$; moreover, it follows from (3.1.14) that

$$\mathbf{E}[\mathrm{d}M_t|\mathcal{F}_t] = 0, \qquad 0 \le t < T. \tag{3.1.17}$$

Thus, our diffusion process $\boldsymbol{X}$ admits a decomposition of the form

$$X_t = X_0 + \int_0^t f(X_s)\,\mathrm{d}s + M_t, \qquad t \in [0,T] \tag{3.1.18}$$

where $\boldsymbol{M} = (M_t)_{t\in[0,T]}$ is a martingale. But, in fact, we can say more about $\boldsymbol{M}$.

First of all, since $t \mapsto f(X_t)$ is continuous by hypothesis, M_t has continuous sample paths. Moreover, $M_0 = 0$, and

$$\begin{aligned}
\mathbf{E}[(\mathrm{d}M_t)^2|\mathcal{F}_t] &= \mathbf{E}[(\mathrm{d}X_t - f(X_t)\,\mathrm{d}t)^2|\mathcal{F}_t] && (3.1.19)\\
&= \mathbf{E}[(\mathrm{d}X_t)^2|\mathcal{F}_t] - 2\mathbf{E}[f(X_t)\,\mathrm{d}X_t|\mathcal{F}_t]\,\mathrm{d}t + \mathbf{E}[f^2(X_t)|\mathcal{F}_t](\mathrm{d}t)^2 && (3.1.20)\\
&= a(X_t)\,\mathrm{d}t - 2f^2(X_t)(\mathrm{d}t)^2 && (3.1.21)\\
&= a(X_t)\,\mathrm{d}t, && (3.1.22)
\end{aligned}$$

which can be rewritten in integral form as

$$\mathbf{E}[(M_t - M_s)^2|\mathcal{F}_s] = \mathbf{E}\left[\int_r^t a(X_r)\,\mathrm{d}r \middle| \mathcal{F}_s\right], \qquad t \ge s. \tag{3.1.23}$$

This readily gives us the quadratic variation of M_t: $\mathrm{d}[M]_t = a(X_t)\,\mathrm{d}t$ or, in integral form,

$$[M]_t = \int_0^t a(X_s)\,\mathrm{d}s, \qquad 0 \le t \le T. \tag{3.1.24}$$

The interesting observation here is that the square of the differential $\mathrm{d}M_t$ is not negligible since it is first order in $\mathrm{d}t$. Thus, we have obtained our first key result:

Theorem 4. *Let $\boldsymbol{X} = (X_t)_{t\in[0,T]}$ be a one-dimensional diffusion process whose drift f and diffusion coefficient a are continuous functions of their argument. Then there exists a square-integrable martingale $\boldsymbol{M} = (M_t)_{t\in[0,T]}$ with continuous sample paths, such that $M_0 = 0$ and*

$$X_t = X_0 + \int_0^t f(X_s)\,\mathrm{d}s + M_t, \qquad 0 \le t \le T. \tag{3.1.25}$$

Moreover, the quadratic variation process $[M]_t$ has the formm

$$[M]_t = \int_0^t a(X_s)\,\mathrm{d}s. \tag{3.1.26}$$

We can consider two special cases: If $a(x) \equiv 0$, then the evolution of X_t is deterministic, i.e.,

$$X_t = X_0 + \int_0^t f(X_s)\,\mathrm{d}s, \qquad 0 \le t \le T \tag{3.1.27}$$

which is equivalent to the ODE

$$\frac{\mathrm{d}}{\mathrm{d}t} X_t = f(X_t), \qquad t \in [0,T] \tag{3.1.28}$$

with (a possibly random) initial condition X_0. On the other hand, if $a(x) = \sigma^2$ for a positive constant $\sigma > 0$, then

$$\mathrm{d}[M]_t = \mathbf{E}[(\mathrm{d}M_t)^2 | \mathcal{F}_t] = \sigma^2\,\mathrm{d}t. \tag{3.1.29}$$

Since $\boldsymbol{X}$ has continuous sample paths and the drift f is a continuous function, it follows that $\boldsymbol{M}$ is a martingale w.r.t. $(\mathcal{F}_t)_{t\in[0,T]}$ that has continuous sample paths. Since $M_0 = 0$, we conclude that the normalized process $(\sigma^{-1}M_t)_{t\ge 0}$ is a *Brownian motion* adapted to the filtration $(\mathcal{F}_t)_{t\ge 0}$ by Lévy's characterization of the Brownian motion (Theorem 3). In other words, we can now express $\boldsymbol{X}$ as a deterministic function of the initial condition X_0 and a Brownian motion $\boldsymbol{W}$:

$$X_t = X_0 + \int_0^t f(X_s)\,\mathrm{d}s + \sigma W_t, \qquad 0 \le t \le T. \tag{3.1.30}$$

As advertised, this expresses $\boldsymbol{X}$ as a deterministic function of X_0 and an independent Brownian motion $\boldsymbol{W}$, and the path $(X_s)_{s\in[0,t]}$ depends only on X_0 and on $(W_s)_{s\in[0,t]}$. Moreover, the Brownian motion $\boldsymbol{W}$ is adapted to the filtration $\mathcal{F}_t$ generated by $\boldsymbol{X}$. We still have to consider the case when the diffusion coefficient a depends on the 'space variable' x.

3.1.3 Enter the stochastic integral

So far, we have been proceeding in a logical fashion, arriving at the general representation of $\boldsymbol{X}$ in Theorem 4, and at the more concrete representation (3.1.30) for $a(x)$ equal to a constant $\sigma^2 > 0$, by deductive reasoning starting from the axioms defining a diffusion process and some basic martingale theory. In order to proceed beyond this simple case, we need to introduce a new tool, that of *stochastic integration*. The basic ideas were developed by Itô (and, somewhat earlier, by Doeblin), where the main point is that one can integrate any sufficiently regular square-integrable process adapted to a filtration w.r.t. a a Brownian motion adapted to the same filtration. However, we will develop the concept of stochastic integration in a bit more generality, which will pay off later when we start talking about filtering and estimation problems.

Let a filtration $(\mathcal{F}_t)_{t\in[0,T]}$ be given, and let $\boldsymbol{M} = (M_t)_{t\in[0,T]}$ be a square-integrable $\mathcal{F}_t$-martingale with continuous sample paths. We would like to be able to define a *stochastic integral*

$$\int_0^T Z_t\,\mathrm{d}M_t \tag{3.1.31}$$

of any process $\boldsymbol{Z} = (Z_t)_{t\in[0,T]}$ adapted to $\mathcal{F}_t$ and satisfying some additional regularity conditions. Any such process will be called the *integrand*, and the martingale $\boldsymbol{M}$ will be called the *integrator*. To that end, let us write the quadratic variation $[M]_t$ in the form $A_t\,\mathrm{d}t$, where A_t is an increasing

positive process adapted to $\mathcal{F}_t$. Then the space of admissible integrands in (3.1.31) will consist of all processes $\boldsymbol{Z} = (Z_t)_{t\in[0,T]}$ that are adapted to $\mathcal{F}_t$ and that also satisfy

$$\int_0^T \mathbf{E}[Z_t^2 A_t]\,\mathrm{d}t < \infty. \tag{3.1.32}$$

We will also use the shorthand $I_T^{\boldsymbol{M}}(\boldsymbol{Z})$ to denote (3.1.31). This is a *definite* integral, and we can obtain the *indefinite* integral $\int_0^t Z\,\mathrm{d}M$ for any $t \in [0,T]$ by setting $\int_0^t Z\,\mathrm{d}M = \int_0^T \mathbf{1}_{[0,t]} Z\,\mathrm{d}M$. Our eventual goal is to show that, when the diffusion coefficient a is everywhere positive, the martingale M_t in (3.1.25) can be expressed as

$$M_t = \int_0^t \sqrt{a(X_s)}\,\mathrm{d}W_s, \qquad 0 \le t \le T \tag{3.1.33}$$

where $\boldsymbol{W}$ is a Brownian motion adapted to $\mathcal{F}_t$. (When a is not everywhere positive, we will still be able to express M_t as an Itô integral, but then we will need to introduce additional 'randomness' to construct the Brownian motion integrator, on top of that already available in $\boldsymbol{X}$.)

We begin by defining the stochastic integral for the so-called *elementary integrands*. We say that an $\mathcal{F}_t$-adapted process $\boldsymbol{Z}$ is an elementary integrand if it can be represented in the form

$$Z_t = \sum_{i=0}^{n-1} \zeta_i \mathbf{1}_{\{t_i \le t < t_{i+1}\}}, \tag{3.1.34}$$

where $t_0 = 0 < t_1 < \ldots < T = t_n$ is a partition of $[0,T]$, and where each ζ_i is a random variable measurable w.r.t. $\mathcal{F}_{t_i}$, such that

$$\sum_{i=0}^{n-1} \mathbf{E}[\zeta_i^2 (M_{t_{i+1}} - M_{t_i})^2] \equiv \int_0^T \mathbf{E}[Z_t^2 A_t]\,\mathrm{d}t < \infty. \tag{3.1.35}$$

We then define the integral (3.1.31) by

$$I_T^{\boldsymbol{M}}(\boldsymbol{Z}) := \sum_{i=0}^{n-1} \zeta_i (M_{t_{i+1}} - M_{t_i}). \tag{3.1.36}$$

Note that, since each ζ_i is $\mathcal{F}_{t_i}$-measurable and $\boldsymbol{M}$ is a martingale, we have

$$\mathbf{E}[\zeta_i (M_{t_{i+1}} - M_{t_i}) | \mathcal{F}_{t_i}] = \zeta_i \mathbf{E}[M_{t_{i+1}} - M_{t_i} | \mathcal{F}_{t_i}] = 0, \tag{3.1.37}$$

so we immediately deduce that $I_T^{\boldsymbol{M}}[\boldsymbol{Z}]$ has zero mean:

$$\mathbf{E}[I_T^{\boldsymbol{M}}(\boldsymbol{Z})] = \mathbf{E}\left[\sum_{i=0}^{n-1} \zeta_i (M_{t_{i+1}} - M_{t_i})\right] \tag{3.1.38}$$

$$= \sum_{i=0}^{n-1} \mathbf{E}[\zeta_i (M_{t_{i+1}} - M_{t_i})] \tag{3.1.39}$$

$$= \sum_{i=0}^{n-1} \mathbf{E}[\mathbf{E}[\zeta_i (M_{t_{i+1}} - M_{t_i})|\mathcal{F}_{t_i}]] \tag{3.1.40}$$

$$= \sum_{i=0}^{n-1} \mathbf{E}[\zeta_i \mathbf{E}[M_{t_{i+1}} - M_{t_i}|\mathcal{F}_{t_i}]] \tag{3.1.41}$$

$$= 0 \tag{3.1.42}$$

Moreover, if $i < j$, then $\mathcal{F}_{t_j} \supseteq \mathcal{F}_{t_{i+1}}$, so

$$\mathbf{E}[\zeta_i \zeta_j (M_{t_{i+1}} - M_{t_i})(M_{t_{j+1}} - M_{t_j})|\mathcal{F}_{t_j}] = \mathbf{E}[\zeta_i \zeta_j (M_{t_{i+1}} - M_{t_i})\mathbf{E}[M_{t_{j+1}} - M_{t_j}]|\mathcal{F}_{t_j}] = 0 \tag{3.1.43}$$

which gives

$$\mathbf{E}\left[(I_T^{\boldsymbol{M}}(\boldsymbol{Z}))^2\right] = \sum_{i,j=0}^{n-1} \mathbf{E}\left[\zeta_i \zeta_j (M_{t_{i+1}} - M_{t_i})(M_{t_{j+1}} - M_{t_j})\right] \tag{3.1.44}$$

$$= \sum_{i=0}^{n-1} \mathbf{E}\left[\zeta_i^2 (M_{t_{i+1}} - M_{t_i})^2\right] \tag{3.1.45}$$

$$= \sum_{i=0}^{n-1} \mathbf{E}\left[\zeta_i^2 \int_{t_i}^{t_{i+1}} A_t \,\mathrm{d}t\right] \tag{3.1.46}$$

$$= \sum_{i=0}^{n-1} \mathbf{E}\left[\int_{t_i}^{t_{i+1}} Z_t^2 A_t \,\mathrm{d}t\right] \tag{3.1.47}$$

$$= \int_0^T \mathbf{E}[Z_t^2 A_t] \,\mathrm{d}t. \tag{3.1.48}$$

One can then prove that, for any admissible integrand $\boldsymbol{Z}$, there exists a sequence $(\boldsymbol{Z}^m)_{m\in\mathbb{N}}$ of elementary integrands, such that

$$\lim_{m\to\infty} \int_0^T \mathbf{E}[(Z_t^m - Z_t)^2 A_t] \,\mathrm{d}t = 0, \tag{3.1.49}$$

and then one can define the stochastic integral as

$$\int_0^T Z_t \,\mathrm{d}M_t := \lim_{m\to\infty} \int_0^T Z_t^m \,\mathrm{d}M_t, \tag{3.1.50}$$

where the limit is in mean-square sense. The detailed proofs can be found in many places, but the main idea is that we can introduce on the space $\mathcal{H}_0$ of elementary integrands an inner product

$$\langle \boldsymbol{Z}, \boldsymbol{Z}' \rangle := \int_0^T \mathbf{E}[Z_t Z_t' A_t] \,\mathrm{d}t, \tag{3.1.51}$$

which is positive definite by vritue of our assumption that $A_t > 0$ for all t, and then consider the Hilbert space $\mathcal{H}$ obtained by completing $\mathcal{H}_0$ w.r.t. the norm $\|\boldsymbol{Z}\| := \sqrt{\langle \boldsymbol{Z}, \boldsymbol{Z} \rangle}$. It can then be shown that $\mathcal{H}$ is precisely the space of all admissible integrands. The stochastic integral has a number of useful properties:

1. Linearity – $\int_0^T (Z_t + Z'_t)\,\mathrm{d}M_t = \int_0^T Z_t\,\mathrm{d}M_t + \int_0^T Z'_t\,\mathrm{d}M_t$
2. Isometry – $\mathbf{E}[(\int_0^T Z_t\,\mathrm{d}M_t)(\int_0^T Z'_t\,\mathrm{d}M_t)] = \int_0^T \mathbf{E}[Z_t Z'_t A_t]\,\mathrm{d}t$
3. Martingale property – the process $(I_t^{\boldsymbol{M}}(\boldsymbol{Z}))_{t\in[0,T]}$ is an $\mathcal{F}_t$-martingale with continuous sample paths

Here, Z_t and Z'_t are arbitrary admissible integrands. All of these are readily verified when Z_t and Z'_t are elementary integrands, and then for arbitrary admissible integrands by taking limits.

Remark 2. *The class of admissible integrands can be expanded even further by relaxing* (3.1.32) *to the weaker requirement*

$$\int_0^T Z_t^2 A_t\,\mathrm{d}t < \infty \quad a.s. \tag{3.1.52}$$

The integral of such a Z_t w.r.t. M_t is then defined as

$$\int_0^T Z_t\,\mathrm{d}M_t := \lim_{n\to\infty} \int_0^T Z_t^n\,\mathrm{d}M_t \text{ in probability} \tag{3.1.53}$$

where the process Z_t^n is defined by

$$Z_t^n := \begin{cases} Z_t, & \text{if } \int_0^t Z_s^2 A_s\,\mathrm{d}s \le n \\ 0, & \text{otherwise} \end{cases} \tag{3.1.54}$$

There is a slight price to pay for this increased generality: The indefinite integral $\int_0^t Z\,\mathrm{d}M$ may no longer be a martingale, but rather a local martingale. *Add informal discussion later.*

3.1.4 Back to diffusion processes

Now we are ready to return to the matter of characterizing the martingale M_t in (3.1.25) under the assumption that the diffusion coefficient a of $\boldsymbol{X}$ is everywhere positive. Recall that the quadratic variation process $[M]_t$ satisfies $\mathrm{d}[M]_t = a(X_t)\,\mathrm{d}t$. It is then readily seen that $Z_t := 1/\sqrt{a(X_t)}$ is an admissible integrand for M_t:

$$\int_0^T \mathbf{E}[Z_t^2 a(X_t)]\,\mathrm{d}t = T. \tag{3.1.55}$$

Let us then define the process W_t by setting $W_0 = 0$ and $\mathrm{d}W_t = \frac{1}{\sqrt{a(X_t)}}\,\mathrm{d}M_t$, or, in integral form,

$$W_t = I_t^{\boldsymbol{M}}(\boldsymbol{Z}) = \int_0^t \frac{1}{\sqrt{a(X_\tau)}}\,\mathrm{d}M_\tau. \tag{3.1.56}$$

This process is an $\mathcal{F}_t$-martingale with continuous sample paths, and we can compute

$$\mathbf{E}[(\mathrm{d}W_t)^2|\mathcal{F}_t] = \frac{1}{a(X_t)}\mathbf{E}[(\mathrm{d}M_t)^2|\mathcal{F}_t] \tag{3.1.57}$$

$$= \frac{1}{a(X_t)}\,\mathrm{d}[M]_t \tag{3.1.58}$$

$$= \frac{1}{a(X_t)}\cdot a(X_t)\,\mathrm{d}t \tag{3.1.59}$$

$$= \mathrm{d}t, \tag{3.1.60}$$

which is equivalent to $\mathrm{d}[W]_t = \mathrm{d}t$. Consequently, $\boldsymbol{W}$ is a *Brownian motion* adapted to $\mathcal{F}_t$ by Lévy's theorem. Writing $\mathrm{d}M_t = \sqrt{a(X_t)}\,\mathrm{d}W_t$, we obtain the following refinement of Theorem 4 and the main result of this section:

Theorem 5 (Doob)**.** *Let $\boldsymbol{X} = (X_t)_{t\in[0,T]}$ be a one-dimensional diffusion process whose drift f and diffusion coefficient a are continuous functions of their argument, and a is everywhere positive. Then there exists a Brownian motion $\boldsymbol{W}$ adapted to $\mathcal{F}_t$, such that*

$$X_t = X_0 + \int_0^t f(X_s)\,\mathrm{d}s + \int_0^t \sqrt{a(X_s)}\,\mathrm{d}W_s. \tag{3.1.61}$$

Remark 3. *If a is only nonnegative, then we have to introduce additional independent randomness into our internalist model. If $a(X_t)$ is positive a.e., then our original construction still goes through. If not, then we define the function*

$$g(x) := \begin{cases} 1/\sqrt{a(x)}, & a(x) > 0 \\ 0, & \textit{otherwise} \end{cases} \tag{3.1.62}$$

and define the process $\boldsymbol{W}$ by $W_0 = 0$ and

$$\mathrm{d}W_t = g(X_t)\,\mathrm{d}M_t + (1 - g(X_t)\sqrt{a(X_t)})\,\mathrm{d}\tilde{W}_t, \tag{3.1.63}$$

where $\tilde{\boldsymbol{W}} = (\tilde{W}_t)_{t\in[0,T]}$ is a Brownian motion independent of $\boldsymbol{X}$. This procedure requires enlarging the filtration $\mathcal{F}_t$ to $\mathcal{G}_t$ containing, in addition to all events generated by $(X_s)_{0\le s\le t}$, also all events generated by $(\tilde{W}_s)_{0\le s\le t}$. It follows from Lévy's theorem that $\boldsymbol{W}$ is a $\mathcal{G}_t$-Brownian motion, and then X_t can be expressed as (3.1.61).

The stochastic integral in (3.1.61), with the Brownian motion serving as the integrator, is the classic stochastic integral of Itô. We will limit ourselves to this setting for the time being, although later we will have many occasions to use more general integrator processes. The main point here is that the process

$$M_t = X_t - X_0 - \int_0^t f(X_s)\,\mathrm{d}s \tag{3.1.64}$$

is a martingale that admits an explicit representation as a stochastic integral w.r.t. a Brownian motion, i.e.,

$$M_t = \int_0^t \sqrt{a(X_s)}\,\mathrm{d}W_s, \tag{3.1.65}$$

and also is an innovations process for X_t.

3.1.5 Multidimensional diffusion processes

The extension to a d-dimensional diffusion process $(X_t)_{t\in[0,T]}$ with drift $f : \mathbb{R}^d \to \mathbb{R}^d$ and diffusion matrix $a : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is straightforward. As before, $\mathcal{F}_t$ denotes the σ-algebra generated by $(X_s)_{s\in[0,t]}$. A d-dimensional Brownian motion $\boldsymbol{W} = (W_t)_{t\in[0,T]}$ is a vector-valued process, $W_t = (W_t^1, \dots, W_t^d)^T$, where W_t^i are independent scalar Brownian motions adapted to $\mathcal{F}_t$.

Theorem 6. *Let $\boldsymbol{X} = (X_t)_{t\in[0,T]}$ be a d-dimensional diffusion process whose drift $f : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ and diffusion matrix $a : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ are continuous functions of their argument, and a is everywhere positive definite. Then there exists a d-dimensional Brownian motion $\boldsymbol{W}$ a adapted to $\mathcal{F}_t$, such that*

$$X_t = X_0 + \int_0^t f(X_s)\,\mathrm{d}s + \int_0^t g(X_s)\,\mathrm{d}W_s, \tag{3.1.66}$$

where $g(x) \in \mathbb{R}^{d\times d}$ is the positive square root of $a(x)$, and the above formula holds coordinatewise, i.e., for each $i \in \{1, \dots, d\}$,

$$X_t^i = X_0^i + \int_0^t f^i(X_s)\,\mathrm{d}s + \sum_{j=1}^d \int_0^t g^{ij}(X_s)\,\mathrm{d}W_s^j, \tag{3.1.67}$$

where $f^i(x)$ is the ith coordinate of $f(x)$ and $g^{ij}(x)$ is the (i, j) entry of $g(x)$.

3.2 Stochastic integration with respect to a Brownian motion

We will now take a closer look at stochastic integrals w.r.t. a Brownian motion. Let a probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in\mathcal{T}}, \mathbf{P})$ be given, where we have explicitly indicated the filtration $(\mathcal{F}_t)_{t\in\mathcal{T}}$. The set $\mathcal{T}$ will be either an interval $[0, T]$ for a finite $T > 0$ or the set of all $t \geq 0$. Let W_t be a $\mathcal{F}_t$-Brownian motion. Since $[W]_t = t$, the space of admissible integrands will consist of all processes $(Z_t)_{t\in\mathcal{T}}$ that are adapted to $(\mathcal{F}_t)_{t\in\mathcal{T}}$ and satisfy

$$\int_{\mathcal{T}} \mathbf{E}[Z_t^2]\,\mathrm{d}t < \infty. \tag{3.2.1}$$

We will refer to the integrals w.r.t. W_t as Itô integrals. The indefinite integral $\int_0^t Z_s\,\mathrm{d}W_s$ is a zero-mean martingale with continuous sample paths, and we the important property of Itô isometry:

$$\mathbf{E}\left[\int_0^t Z_s\,\mathrm{d}W_s\right] = \int_0^t \mathbf{E}[Z_s^2]\,\mathrm{d}s. \tag{3.2.2}$$

In fact, as discussed in Remark 2, we can consider a wider class of admissible integrands Z_t that are $\mathcal{F}_t$-adapted and satisfy

$$\int_{\mathcal{T}} Z_t^2\,\mathrm{d}t < \infty \text{ a.s.} \tag{3.2.3}$$

We will use this version from now on since it will allow us to state various results to a useful degree of generality, even though the resulting indefinite integrals may no longer be martingales.

It helps to see an example to appreciate how different Itô integration is compared to the rules of ordinary integral calculus. For example, let's take $Z_t = W_t$, which is clearly admissible on any finite interval $\mathcal{T} = [0, T]$. It is tempting to conjecture, by analogy with ordinary calculus, that

$$\int_0^t W_s \, \mathrm{d}W_s = \frac{1}{2} W_t^2. \tag{3.2.4}$$

This, however, cannot be the case since $\int_0^t W_s \, \mathrm{d}W_s$ must have zero mean, while $\mathbf{E}[W_t^2] = t$. In order to compute the Itô integral of W_t w.r.t. $\mathrm{d}W_t$, let us consider any sequence of increasingly fine partitions of $[0, t]$. For any such partition $(t_k)_{k=0}^m$ with $t_0 = 0$ and $t_m = t$, we can write

$$\frac{1}{2} W_t^2 = \sum_{k=0}^{m-1} W_{t_k} (W_{t_{k+1}} - W_{t_k}) + \frac{1}{2} \sum_{k=0}^{m-1} (W_{t_{k+1}} - W_{t_k})^2. \tag{3.2.5}$$

Then, taking the limit of both sides as the partitions get finer and finer, we see that the first term on the right-hand side converges the integral $\int_0^t W_s \, \mathrm{d}W_s$, while the second term converges to $\frac{1}{2}[W]_t = \frac{1}{2} t$, so

$$\int_0^t W_s \, \mathrm{d}W_s = \frac{1}{2} (W_t^2 - t). \tag{3.2.6}$$

However, we have derived this using a bespoke argument that can't be generalized further. In order to have a general rule that will allow us to compute Itô integrals, we need to introduce stochastic differentials and Itô's differentiation rule.

3.2.1 Stochastic differentials and Itô's differentiation rule

Let F_t and G_t be $\mathcal{F}_t$-adapted processes, such that

$$\int_{\mathcal{T}} |F_t| \, \mathrm{d}t < \infty \text{ and } \int_{\mathcal{T}} G_t^2 \, \mathrm{d}t < \infty \quad \text{a.s.} \tag{3.2.7}$$

and let X_0 be an $\mathcal{F}_0$-measurable random variable. Then we can form the *integral process*

$$X_t := X_0 + \int_0^t F_s \, \mathrm{d}s + \int_0^t G_s \, \mathrm{d}W_s, \qquad t \in \mathcal{T} \tag{3.2.8}$$

where the second term on the right-hand side is an ordinary integral, while the third term is an Itô integral. We can also write this as an *Itô stochastic differential*

$$\mathrm{d}X_t = F_t \, \mathrm{d}t + G_t \, \mathrm{d}W_t, \tag{3.2.9}$$

which, together with the initial condition X_0, is always to be interpreted as shorthand for (3.2.8). Such processes are also referred to as *Itô processes*. We can integrate other processes w.r.t. X_t: If V_t is an $\mathcal{F}_t$-adapted process, such that

$$\int_{\mathcal{T}} |V_t F_t| \, \mathrm{d}t < \infty \text{ and } \int_{\mathcal{T}} V_t^2 G_t^2 \, \mathrm{d}t < \infty \quad \text{a.s.}, \tag{3.2.10}$$

then we can define the integral of V_t w.r.t. X_t in an obvious fashion:

$$\int_0^t V_s \, \mathrm{d}X_s := \int_0^t V_s F_s \, \mathrm{d}s + \int_0^t V_s G_s \, \mathrm{d}W_s. \tag{3.2.11}$$

This gives another Itô process. It is not hard to see that the class of all Itô processes is closed under linear operations: If X_t and $\tilde{X}_t$ are two Itô processes, then any linear combination $\alpha X_t + \beta \tilde{X}_t$, where α and β are deterministic constants, is also an Itô process. In fact, as we will see shortly, the class of Itô processes is closed under multiplication, i.e., $Y_t = X_t \tilde{X}_t$ will also be an Itô process. This is a consequence of the fact that any sufficiently smooth functional of an Itô process is itself an Itô process. We begin by stating a simple one-dimensional version:

Theorem 7 (Itô's differentiation rule, one-dimensional version)**.** *Let X_t be an integral process, as in (3.2.8), and consider the process $Y_t := \varphi(X_t, t)$, where $\varphi(x,t)$ is once continuously differentiable w.r.t. t and twice continuously differentiable w.r.t. x. Then the following holds almost surely:*

$$Y_t = \varphi(X_0, 0) + \int_0^t \dot{\varphi}(X_s, s) \, \mathrm{d}s + \int_0^t \varphi'(X_s, s) \, \mathrm{d}X_s + \frac{1}{2} \int_0^t \varphi''(X_s, s) F_s \, \mathrm{d}s, \tag{3.2.12}$$

where $\dot{\varphi}(x,t) := \frac{\partial}{\partial t}\varphi(x,t)$, $\varphi'(x,t) := \frac{\partial}{\partial x}\varphi(x,t)$, and $\varphi''(x,t) := \frac{\partial^2}{\partial x^2}\varphi(x,t)$.

Proof. We will give a formal argument instead of a rigorous proof to illustrate the main idea. We are interested in computing the forward differential

$$\mathrm{d}Y_t = Y_{t+\mathrm{d}t} - Y_t = \varphi(X_{t+\mathrm{d}t}, t + \mathrm{d}t) - \varphi(X_t, t). \tag{3.2.13}$$

Since $\varphi(x,t)$ is twice differentiable in x and once in t, we can write out a Taylor expansion of φ to second order x and to first order in t:

$$\begin{aligned}
&\varphi(X_{t+\mathrm{d}t}, t + \mathrm{d}t) - \varphi(X_t, t) && (3.2.14)\\
&= \varphi(X_t + \mathrm{d}X_t, t + \mathrm{d}t) - \varphi(X_t, t) && (3.2.15)\\
&= \varphi'(X_t, t) \, \mathrm{d}X_t + \dot{\varphi}(X_t, t) \, \mathrm{d}t + \frac{1}{2}\varphi''(X_t, t)(\mathrm{d}X_t)^2 + \text{higher-order terms} && (3.2.16)
\end{aligned}$$

Since $\mathrm{d}X_t = F_t \, \mathrm{d}t + G_t \, \mathrm{d}W_t$, we can express $(\mathrm{d}X_t)^2$ further as

$$\begin{aligned}
(\mathrm{d}X_t)^2 &= F_t^2 (\mathrm{d}t)^2 + 2F_t G_t \, \mathrm{d}W_t \, \mathrm{d}t + G_t^2 (\mathrm{d}W_t)^2 && (3.2.17)\\
&= F_t^2 (\mathrm{d}t)^2 + 2F_t G_t \, \mathrm{d}W_t \, \mathrm{d}t + G_t^2 \, \mathrm{d}t, && (3.2.18)
\end{aligned}$$

where the equality $(\mathrm{d}W_t)^2 = \mathrm{d}t$ holds almost surely and in mean-square sense. All the other terms are $o(\mathrm{d}t)$, so we can neglect them and write

$$(\mathrm{d}X_t)^2 = G_t^2 \, \mathrm{d}t. \tag{3.2.19}$$

Substituting this back into our expression for $\mathrm{d}Y_t$ gives

$$\mathrm{d}Y_t = \dot{\varphi}(X_t, t) \, \mathrm{d}t + \varphi'(X_t, t) \, \mathrm{d}X_t + \frac{1}{2}\varphi''(X_t) G_t^2 \, \mathrm{d}t \tag{3.2.20}$$

and (3.2.12) follows by integration. □

Thus, Itô's differentiation rule (or, simply, Itô's rule) tells us that, for any function $\varphi(x,t)$ which is C^2 in x and C^1 in t and for any Itô process $\mathrm{d}X_t = F_t\,\mathrm{d}t + G_t\,\mathrm{d}W_t$, $Y_t = \varphi(X_t,t)$ is itself an Itô process, i.e.,

$$\mathrm{d}Y_t = \left(\dot{\varphi}(X_t,t) + \varphi'(X_t,t)F_t + \frac{1}{2}\varphi''(X_t)G_t^2\right)\mathrm{d}t + \varphi'(X_t,t)G_t\,\mathrm{d}W_t. \tag{3.2.21}$$

The presence of the last term on the right-hand side of (3.2.12), involving the second derivative of $\varphi(x,t)$ w.r.t. x, is the surprising part and the main difference between the rules of Itô calculus and those of ordinary calculus. In order to appreciate how different this really is, consider the case when $G_t \equiv 0$ for all t. Then, on the path-by-path basis, we have the ODE

$$\frac{\mathrm{d}}{\mathrm{d}t}X_t(\omega) = F_t(\omega), \tag{3.2.22}$$

and we can simply apply the chain rule to see what is happening with $Y_t = \varphi(X_t,t)$:

$$\frac{\mathrm{d}}{\mathrm{d}t}Y_t(\omega) = \frac{\mathrm{d}}{\mathrm{d}t}\varphi(X_t(\omega),t) \tag{3.2.23}$$

$$= \dot{\varphi}(X_t(\omega),t) + \varphi'(X_t(\omega),t)\frac{\mathrm{d}}{\mathrm{d}t}X_t(\omega), \tag{3.2.24}$$

which we can also write as $\mathrm{d}Y_t = \dot{\varphi}(X_t,t)\,\mathrm{d}t + \varphi'(X_t,t)\,\mathrm{d}X_t$. Integrating, we obtain the expression

$$\varphi(X_t,t) = \varphi(X_0,0) + \int_0^t \dot{\varphi}(X_s,s)\,\mathrm{d}s + \int_0^t \varphi'(X_s)\,\mathrm{d}X_s, \tag{3.2.25}$$

which is exactly what we would expect from the rules of ordinary calculus. The extra term in Itô's rule that contains φ'' is a consequence of the fact that $\mathrm{d}[W]_t = \mathrm{d}t$. Ultimately, we will need the multidimensional version of Itô's rule, but we can already look at a few examples to get an idea of what's involved.

Example 1 ($\int_0^t W\,\mathrm{d}W$ revisited)**.** *The formula* (3.2.6) *for the Itô integral of W w.r.t. $\mathrm{d}W$ can now be seen as a direct consequence of Itô's rule. Indeed, since W_t is evidently an Itô process with $F_t = 0$ and $G_t = 1$, we can apply Theorem 7 to $\varphi(x,t) = \frac{1}{2}x^2$. Then*

$$\dot{\varphi}(x,t) = 0, \quad \varphi'(x,t) = x, \quad \varphi''(x,t) = 1 \tag{3.2.26}$$

and therefore

$$\frac{1}{2}W_t^2 = \int_0^t W_s\,\mathrm{d}W_s + \frac{1}{2}\int_0^t \mathrm{d}s = \int_0^t W_s\,\mathrm{d}W_s + \frac{1}{2}t. \tag{3.2.27}$$

Rearranging, we get (3.2.6).

Example 2 (fundamental theorem of Itô calculus)**.** *We can generalize the above example to compute the Itô integral $\int_0^t \varphi'(W_t)\,\mathrm{d}W_t$, where $\varphi : \mathbb{R} \to \mathbb{R}$ is a twice continuously differentiable function. In this case, Itô's rule gives the formula*

$$\varphi(W_t) = \varphi(0) + \int_0^t \varphi'(W_s)\,\mathrm{d}W_s + \frac{1}{2}\int_0^t \varphi''(W_s)\,\mathrm{d}s, \tag{3.2.28}$$

or, upon rearranging,

$$\int_0^t \varphi'(W_s)\,\mathrm{d}W_s = \varphi(W_t) - \varphi(0) - \frac{1}{2}\int_0^t \varphi''(W_s)\,\mathrm{d}s. \tag{3.2.29}$$

We can also consider functions $\varphi(x,t)$. For example, let $\varphi(x,t) = e^{x-t/2}$. Then

$$\dot{\varphi}(x,t) = -\frac{1}{2}e^{x-t/2} = -\frac{1}{2}\varphi(x,t), \quad \varphi'(x,t) = \varphi''(x,t) = \varphi(x,t). \tag{3.2.30}$$

Applying Itô's rule gives

$$e^{W_t - t/2} = 1 - \frac{1}{2}\int_0^t e^{W_s - s/2}\,\mathrm{d}s + \int_0^t e^{W_s - s/2}\,\mathrm{d}W_s + \frac{1}{2}\int_0^t e^{W_s - s/2}\,\mathrm{d}s \tag{3.2.31}$$

$$1 + \int_0^t e^{W_s - s/2}\,\mathrm{d}W_s \tag{3.2.32}$$

so, for this particular choice of φ, we can write

$$\int_0^t \varphi(W_s, s)\,\mathrm{d}W_s = \varphi(W_s, s)\big|_0^t. \tag{3.2.33}$$

Compare this with ordinary calculus, where the exponential function $x \mapsto e^x$ is the unique solution of the ODE $\varphi'(x) = \varphi(x)$, i.e.,

$$\int_a^b \varphi(x)\,\mathrm{d}x = \varphi(x)\big|_a^b. \tag{3.2.34}$$

We will later see that this is a special case of the so-called Doléans-Dade exponential *of an Itô process.*

3.2.2 Multidimensional Itô calculus

Since many applications of interest involve multidimensional processes, we need to develop the appropriate framework for working with such processes. As before, we have the probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in\mathcal{T}}, \mathbf{P})$ with a filtration defined on it, as well as an m-dimensional Brownian motion process $W_t = (W_t^1, \ldots, W_t^m)^T$ adapted to this filtration. (That is, W_t^i are independent Brownian motions adapted to $\mathcal{F}_t$.) We then say that $X_t = (X_t^1, \ldots, X_t^n)^T$ is an n-dimensional Itô process adapted to $\mathcal{F}_t$ if there exist adapted processes F_t^i, G_t^{ij} for $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$, such that

$$\int_{\mathcal{T}} |F_t^i|\,\mathrm{d}t < \infty \text{ and } \int_{\mathcal{T}} |G_t^{ij}|^2\,\mathrm{d}t < \infty \text{ a.s.} \qquad \forall i, j \tag{3.2.35}$$

and

$$X_t^i = X_0^i + \int_0^t F_s^i\,\mathrm{d}s + \sum_{j=1}^m \int_0^t G_s^{ij}\,\mathrm{d}W_s^j, \qquad i = 1, \ldots, n. \tag{3.2.36}$$

If we assemble the processes F_t^i into an n-dimensional vector process $F_t = (F_t^1, \ldots, F_t^n)^T$ and G_t^{ij} into an $n \times m$ matrix process G_t, then we can express (3.2.36) more succinctly as

$$X_t = X_0 + \int_0^t F_s\,\mathrm{d}s + \int_0^t G_s\,\mathrm{d}W_s. \tag{3.2.37}$$

We can now state the multidimensional version of Itô's rule:

Theorem 8. *Let $\varphi : \mathbb{R}^n \times \mathcal{T} \to \mathbb{R}$ be such that $\varphi(x,t)$, where $x = (x^1, \dots, x^n)^T$, is twice continuously differentiable in all x^i and once continuously differentiable in t. Then $Y_t = \varphi(X_t, t)$ is an Itô process:*

$$\varphi(X_t,t) = \varphi(X_0,0) + \sum_{i=1}^{n}\sum_{j=1}^{m}\int_0^t \varphi_i(X_s,s)G_s^{ij}\,\mathrm{d}W_s^j + \int_0^t \left(\dot\varphi(X_s,s) + \sum_{i=1}^{n}\varphi_i(X_s,s)F_s^i + \frac{1}{2}\sum_{i,j=1}^{n}\sum_{k=1}^{m}\varphi_{ij}(X_s,s)G_s^{ik}G_s^{jk}\right)\mathrm{d}s,$$

where $\dot\varphi(x,s) := \frac{\partial}{\partial s}\varphi(x,s)$, $\varphi_i(x,s) := \frac{\partial}{\partial x^i}\varphi(x,s)$, $\varphi_{ij}(x,s) := \frac{\partial^2}{\partial x^i \partial x^j}\varphi(x,s)$.

We can rewrite (3.2.38) using a more compact vector notation as

$$\mathrm{d}\varphi(X_t,t) = \dot\varphi(X_t,t)\,\mathrm{d}t + \partial\varphi(X_t,t)\,\mathrm{d}X_t + \frac{1}{2}\mathrm{tr}\left\{\partial^2\varphi(X_t,t)\,\mathrm{d}X_t(\mathrm{d}X_t)^T\right\}, \tag{3.2.38}$$

where $\partial\varphi(x,t)$ denotes the row vector $(\varphi_1(x,t), \dots, \varphi_n(x,t))$, $\partial^2\varphi(x,t)$ is the $n \times n$ matrix with entries $\varphi_{ij}(x,t)$, and the products $\mathrm{d}X_t^i\,\mathrm{d}X_t^j$ are evaluated using the *box rules*

$$(\mathrm{d}t)^2 = \mathrm{d}W_t^i\,\mathrm{d}t = \mathrm{d}t\,\mathrm{d}W_t^i = 0, \qquad \mathrm{d}W_t^i\,\mathrm{d}W_t^j = \delta_{ij}\,\mathrm{d}t. \tag{3.2.39}$$

Example 3 (Itô product rule)**.** *Let X_t^i, $i \in \{1,2\}$, be two scalar Itô processes. Then, for the function $\varphi(x^1,x^2) = x^1x^2$ we have*

$$\partial\varphi(x^1,x^2) = (x^2,x^1), \quad \partial^2\varphi(x^1,x^2) = \begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix} \tag{3.2.40}$$

so Itô's rule gives

$$\mathrm{d}(X_t^1X_t^2) = (X_t^2,X_t^1)\begin{pmatrix}\mathrm{d}X_t^1\\ \mathrm{d}X_t^2\end{pmatrix} + \frac{1}{2}(\mathrm{d}X_t^1,\mathrm{d}X_t^2)\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}\begin{pmatrix}\mathrm{d}X_t^1\\ \mathrm{d}X_t^2\end{pmatrix} \tag{3.2.41}$$

$$= X_t^1\,\mathrm{d}X_t^2 + X_t^2\,\mathrm{d}X_t^1 + \mathrm{d}X_t^1\,\mathrm{d}X_t^2, \tag{3.2.42}$$

which is the Itô calculus counterpart of the product rule $\mathrm{d}(UV) = U\,\mathrm{d}V + V\,\mathrm{d}U$ of ordinary calculus, where we pick up the extra term $\mathrm{d}U\,\mathrm{d}V$. Among other things, this shows that the class of Itô processes is closed under pontwise multiplication – this is easy to verify by starting with the representations $\mathrm{d}X_t^i = F_t^i\,\mathrm{d}t + G_t^i\,\mathrm{d}W_t$, multiplying things out, and using the box rules.

The product rule can be written down in a more compact form in terms of the joint quadratic variation *of X^1 and X^2, which is denoted by $[X^1,X^2]$ and defined as*

$$[X^1,X^2] := \frac{1}{4}\Big([X^1+X^2] - [X^1-X^2]\Big). \tag{3.2.43}$$

Using this definition and Itô's box rules, it is not hard to show that

$$\mathrm{d}X_t^1\,\mathrm{d}X_t^2 = G_t^1G_t^2\,\mathrm{d}t = \mathrm{d}[X^1,X^2]_t, \tag{3.2.44}$$

so the product rule now takes the form

$$\mathrm{d}(X_t^1X_t^2) = X_t^1\,\mathrm{d}X_t^2 + X_t^2\,\mathrm{d}X_t^1 + \mathrm{d}[X^1,X^2]_t. \tag{3.2.45}$$

Example 4 (Proof of Lévy's theorem)**.** *Let $(W_t)_{t\geq 0}$ be a martingale w.r.t. a filtration $(\mathcal{F})_{t\geq 0}$ with $W_0 = 0$, and suppose that W_t has continuous sample paths and $[W]_t = t$. Lévy's characterization of Brownian motion (Theorem 3) then says that W_t is a Brownian motion w.r.t. $\mathcal{F}_t$. To see this, let us apply Itô's lemma to $Z_t = e^{i\alpha W_t}$, where α is an arbitrary real parameter and $i = \sqrt{-1}$ is the imaginary unit: For any $t \geq s \geq 0$,*

$$\mathrm{d}Z_t = i\alpha Z_t \,\mathrm{d}W_t - \frac{1}{2}\alpha^2 Z_t \,\mathrm{d}[W]_t \tag{3.2.46}$$

$$= i\alpha Z_t \,\mathrm{d}W_t - \frac{1}{2}\alpha^2 Z_t \,\mathrm{d}t. \tag{3.2.47}$$

Therefore, for $t \geq s \geq 0$,

$$e^{i\alpha W_t} - e^{i\alpha W_s} = i\alpha \int_s^t Z_r \,\mathrm{d}W_r - \frac{1}{2}\alpha^2 \int_s^t Z_r \,\mathrm{d}r. \tag{3.2.48}$$

Using the martingale property of stochastic integrals gives

$$\mathbf{E}[e^{i\alpha W_t}|\mathcal{F}_s] = e^{i\alpha W_s} - \frac{\alpha^2}{2}\int_s^t \mathbf{E}[e^{i\alpha W_r}|\mathcal{F}_s]\,\mathrm{d}r. \tag{3.2.49}$$

This can be easily solved for $\mathbf{E}[e^{i\alpha W_t}|\mathcal{F}_s]$:

$$\mathbf{E}[e^{i\alpha W_t}|\mathcal{F}_s] = e^{i\alpha W_s} e^{-\frac{\alpha^2}{2}(t-s)}, \tag{3.2.50}$$

which is equivalent to $\mathbf{E}[e^{i\alpha(W_t - W_s)}|\mathcal{F}_s] = e^{-\frac{\alpha^2}{2}(t-s)}$ for $t \geq s \geq 0$. Thus, conditionally on $\mathcal{F}_s$, the increment $W_t - W_s$ is Gaussian with zero mean and variance $t - s$. This proves that $(W_t)_{t\geq 0}$ is an $\mathcal{F}_t$-Brownian motion.

3.3 Stochastic differential equations

We have begun this chapter by presenting Doob's solution of the realization problem for a class of diffusion processes: If $(X_t)_{t\in[0,T]}$ is a d-dimensional diffusion process with drift $f(x)$ and positive definite diffusion matrix $a(x)$, then it can be realized as an Itô process

$$X_t = X_0 + \int_0^t f(X_s)\,\mathrm{d}s + \int_0^t g(X_s)\,\mathrm{d}W_s, \qquad 0 \leq t \leq T \tag{3.3.1}$$

for some d-dimensional Brownian motion adapted to the filtration generated by X_t. Now, Eq. (3.3.1) expresses X_t as an Itô process with $F_t = f(X_t)$ and $G_t = g(X_t)$ both depending on X_t, i.e.,

$$\mathrm{d}X_t = f(X_t)\,\mathrm{d}t + g(X_t)\,\mathrm{d}W_t, \qquad 0 \leq t \leq T \tag{3.3.2}$$

with the given initial condition X_0. Doob's representation theorem then allows us to say that $(X_t)_{t\in[0,T]}$ is a *solution* of the *stochastic differential equation* (SDE) (3.3.2).

This observation motivates us to ask this question generally: Let an m-dimensional Brownian motion $(W_t)_{t\in[0,T]}$, an n-dimensional random vector X_0 independent of (W_t), and functions $f : \mathbb{R}^n \to$

$\mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^{n\times m}$ be given. We would like to know whether there exists a solution of the stochastic differential equation

$$\mathrm{d}X_t = f(X_t)\,\mathrm{d}t + g(X_t)\,\mathrm{d}W_t, \qquad 0 \le t \le T \tag{3.3.3}$$

i.e., an adapted n-dimensional process $(X_t)_{t\in[0,T]}$, such that

$$X_t = X_0 + \int_0^t f(X_s)\,\mathrm{d}s + \int_0^t g(X_s)\,\mathrm{d}W_s, \qquad 0 \le t \le T. \tag{3.3.4}$$

Moreover, we would like to know whether X_t is a diffusion process and to identify its drift and diffusion matrix. Thus, we are starting from an internalist description given by (3.3.3) in terms of the initial condition X_0 and an independent Brownian motion playing the role of the latent variables; we then ask whether this description is well-posed in the sense of admitting a well-behaved externalist description, i.e., a solution of (3.3.3) with the latent Brownian motion eliminated. As we will see next, this is possible under mild regularity conditions on f and g.

A reminder about notation. Given a vector $v \in \mathbb{R}^n$ and a matrix $A \in \mathbb{R}^{n\times m}$, we will denote by $|v|$ and $\|A\|$ the Euclidean norm of v and the operator norm of A, i.e.,

$$|v| = \sqrt{v^T v}, \qquad \|A\| = \sup\{|Au| : u \in \mathbb{R}^m, |u| = 1\}. \tag{3.3.5}$$

Theorem 9. *Suppose that the initial condition X_0 and the functions f, g satisfy the following for some constant $0 \le K < \infty$:*

1. *X_0 is independent of $(W_t)_{t\in[0,T]}$, and $\mathbf{E}[|X_0|^2] < \infty$*
2. *f and g are Lipschitz continuous:*

$$|f(x) - f(y)| + \|g(x) - g(y)\| \le K|x - y|, \qquad x, y \in \mathbb{R}^n \tag{3.3.6}$$

3. *f and g are of at most linear growth:*

$$|f(x)|^2 + \|g(x)\|^2 \le K^2(1 + |x|^2), \qquad x \in \mathbb{R}^n \tag{3.3.7}$$

Then the SDE (3.3.3) has a unique solution $(X_t)_{t\in[0,T]}$ satisfying $\int_0^t \mathbf{E}[|X_s|^2]\,\mathrm{d}s < \infty$ for all $t \in [0, T]$. Moreover, X_t is a diffusion process with drift $f(x)$ and diffusion matrix $a(x) = g(x)g(x)^T$.

Proof. The proof is based on *Picard iteration*, which is the same successive approximation procedure one uses to show existence and uniqueness of solutions of ODEs. To keep things simple, we will limit ourselves to the scalar case, i.e., $n = m = 1$; the general case is proved along the same lines. We will construct a sequence of processes $(X^{(n)})_{n=0,1,\ldots}$ recursively as follows: $X_t^{(0)} = X_0$ for all $t \in [T]$, and for $n = 0, 1, \ldots$

$$X_t^{(n+1)} := X_0 + \int_0^t f(X_s^{(n)})\,\mathrm{d}s + \int_0^t g(X_s^{(n)})\,\mathrm{d}W_s, \qquad 0 \le t \le T \tag{3.3.8}$$

Each of these processes is evidently adapted. Moreover, using the linear growth condition (3.3.7), it can be verified that each $X_t^{(n)}$ has continuous sample paths and satisfies $\int_0^t \mathbf{E}[|X_s^{(n)}|^2]\,\mathrm{d}s < \infty$ for all $0 \le t \le T$. Next, let $\Delta_t^n := \mathbf{E}[(X_t^{(n+1)} - X_t^{(n)})^2]$. Then, since

$$X_t^{(n+1)} - X_t^{(n)} = \int_0^t (f(X_s^{(n)}) - f(X_s^{(n-1)}))\,\mathrm{d}s + \int_0^t (g(X_s^{(n)}) - g(X_s^{(n-1)}))\,\mathrm{d}W_s, \tag{3.3.9}$$

we can estimate

$$\Delta_t^n \le 2\mathbf{E}\left[\left(\int_0^t (f(X_s^{(n)}) - f(X_s^{(n-1)}))\,\mathrm{d}s\right)^2\right] + 2\mathbf{E}\left[\left(\int_0^t (g(X_s^{(n)}) - g(X_s^{(n-1)}))\,\mathrm{d}W_s\right)^2\right] \tag{3.3.10}$$

$$\le 2t\int_0^t \mathbf{E}[(f(X_s^{(n)}) - f(X_s^{(n-1)}))^2]\,\mathrm{d}s + 2\int_0^t \mathbf{E}[(g(X_s^{(n)}) - g(X_s^{(n-1)}))^2]\,\mathrm{d}s \tag{3.3.11}$$

$$\le 2K^2(1+T)\int_0^t \mathbf{E}[|X_s^{(n)} - X_s^{(n-1)}|^2]\,\mathrm{d}s \tag{3.3.12}$$

$$=: C_0\int_0^t \Delta_s^{n-1}\,\mathrm{d}s, \tag{3.3.13}$$

where we have used the Itô isometry and the Lipschitz continuity condition (3.3.6). Now,

$$\Delta_t^0 = \mathbf{E}[(X_t^{(1)} - X_t^{(0)})^2] \tag{3.3.14}$$

$$= \mathbf{E}\left[\left(\int_0^t f(X_0)\,\mathrm{d}s + \int_0^t g(X_0)\,\mathrm{d}W_s\right)^2\right] \tag{3.3.15}$$

$$\le 2\mathbf{E}[t|f(X_0)|^2 + |g(X_0)|^2] \tag{3.3.16}$$

$$\le 2K^2(1+T)\mathbf{E}[|X_0|^2]t \tag{3.3.17}$$

$$=: C_1 t. \tag{3.3.18}$$

Therefore, repeated integration gives the estimate

$$\Delta_t^n \le C_1\frac{(C_0 t)^n}{n!}. \tag{3.3.19}$$

Now, for any $t \in [0,T]$ and any $k, n \in \{0,1,\dots\}$, we can use Cauchy–Schwarz inequality to write

$$|X_t^{(n+k)} - X_t^{(n)}| \le \sum_{j=1}^k |X_t^{(n+j)} - X_t^{(n+j-1)}| \tag{3.3.20}$$

$$\le \left(\sum_{j=1}^k \frac{1}{2^{n+j-1}}\right)^{1/2}\left(\sum_{j=1}^k 2^{n+j-1}|X_t^{(n+j)} - X_t^{(n+j-1)}|^2\right)^{1/2}, \tag{3.3.21}$$

so

$$\mathbf{E}\left[|X_t^{(n+k)} - X_t^{(n)}|^2\right] \le \sum_{j=1}^{k} \frac{1}{2^{n+j-1}} \cdot \sum_{j=1}^{k} 2^{n+j-1}\Delta_t^{n+j-1} \tag{3.3.22}$$

$$\le 2C_1 \sum_{j=1}^{k} \frac{(2C_0 t)^{n+j-1}}{(n+j-1)!} \tag{3.3.23}$$

$$\le 2C_1 \sum_{j=n-1}^{\infty} \frac{(2C_0 t)^j}{j!}, \tag{3.3.24}$$

which means that

$$\sup_{k\ge 0} \mathbf{E}\left[|X_t^{(n+k)} - X_t^{(n)}|^2\right] \le 2C_1 \sum_{j=n-1}^{\infty} \frac{(2C_0 T)^j}{j!}, \tag{3.3.25}$$

where the right-hand side converges to 0 as $n \to \infty$. Thus

$$\sup_{t\in[0,T]} \sup_{k\ge 0} \mathbf{E}\left[|X_t^{(n+k)} - X_t^{(n)}|^2\right] \xrightarrow{n\to\infty} 0. \tag{3.3.26}$$

This shows that, for every $t \in [0,T]$, the sequence $(X_t^{(n)})_{n=0,1,\dots}$ converges in mean-square, so we have obtained a process $(X_t)_{t\in[0,T]}$, such that

$$\lim_{n\to\infty} \sup_{t\in[0,T]} \mathbf{E}\left[|X_t^{(n)} - X_t|^2\right] = 0. \tag{3.3.27}$$

This process is adapted, satisfies $\int_0^t \mathbf{E}[X_s^2]\,\mathrm{d}s < \infty$ for all $t \in [0,T]$, and solves (3.3.3). To show uniqueness, let $(\tilde{X}_t)_{t\in[0,T]}$ be another solution of (3.3.3). Then $Z_t := X_t - \tilde{X}_t$ satisfies $Z_0 = 0$ and

$$Z_t = \int_0^t (f(X_s) - f(\tilde{X}_s))\,\mathrm{d}s + \int_0^t (g(X_s) - g(\tilde{X}_s))\,\mathrm{d}W_s, \tag{3.3.28}$$

and the same argument as before gives

$$\mathbf{E}Z_t^2 \le 2K^2(1+T)\int_0^t \mathbf{E}Z_s^2\,\mathrm{d}s, \qquad 0 \le t \le T \tag{3.3.29}$$

Gronwall's inequality then gives $\mathbf{E}Z_t^2 \le e^{2K^2(1+T)t}\mathbf{E}Z_0^2$, so $\mathbf{E}Z_t^2 = 0$ for all $t \in [0,T]$. Verification of the fact that X_t is a diffusion process with drift f and diffusion matrix gg^T is left as an exercise. □

3.3.1 Examples

Before proceeding further, let us see some examples of SDEs that admit explicit solutions.

Example 5 (Linear SDEs and the Ornstein–Uhlenbeck process)**.** *A (time-homogeneous) linear ODE is of the form*

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = Fx(t), \tag{3.3.30}$$

where $x(t)$ is an n-dimensional vector and A is an $n \times n$ matrix. The solution, for a given initial condition at $x = 0$, is given by the matrix exponential: $x(t) = e^{tF}x(0)$. The stochastic analog takes the form

$$\mathrm{d}X_t = FX_t\,\mathrm{d}t + G\,\mathrm{d}W_t \tag{3.3.31}$$

for an n-dimensional process X_t and an m-dimensional Brownian motion W_t, where $F \in \mathbb{R}^{n\times n}$ and $G \in \mathbb{R}^{n\times m}$ are given matrices. Here, $f(x) = Fx$ and $g(x) = G$ evidently satisfy the regularity conditions of Theorem 9, so for any $T > 0$ there exists a unique diffusion process $(X_t)_{t\in[0,T]}$ with drift $f(x) = Fx$ and diffusion matrix $a(x) = GG^T$ that solves the above SDE for any square-integrable initial condition X_0. We will now show that this process can be expressed in closed form in terms of the initial condition X_0 and the Brownian motion W_t only.

To that end, consider the vector-valued function $\varphi(x, t) = e^{-tF}x$. Applying Itô's rule coordinate-wise, we can write the following for the process $Y_t = \varphi(X_t, t) = e^{-tF}X_t$:

$$\begin{aligned}\mathrm{d}Y_t &= -F\varphi(X_t, t)\,\mathrm{d}t + e^{-tF}\,\mathrm{d}X_t &(3.3.32)\\ &= -F\varphi(X_t, t)\,\mathrm{d}t + e^{-tF}FX_t\,\mathrm{d}t + e^{-tF}G\,\mathrm{d}W_t &(3.3.33)\\ &= e^{-tF}G\,\mathrm{d}W_t, &(3.3.34)\end{aligned}$$

where the term $\frac{1}{2}\,\mathrm{d}X_t(\mathrm{d}X_t)^T\partial^2\varphi(X_t, t)$ vanishes since $\varphi(x, t)$ is linear in x. Since $Y_0 = X_0$, we can integrate the above SDE to get

$$Y_t = X_0 + \int_0^t e^{-sF}G\,\mathrm{d}W_s, \qquad 0 \le t \le T \tag{3.3.35}$$

which finally gives

$$X_t = e^{tF}X_0 + \int_0^t e^{(t-s)F}G\,\mathrm{d}W_s. \tag{3.3.36}$$

This process is known as the Ornstein–Uhlenbeck process*. If X_0 is a Gaussian random vector, then X_t is a Gaussian process.*

Example 6 (Stochastic exponential of Brownian motion)**.** *Consider the one-dimensional SDE*

$$\mathrm{d}X_t = X_t\,\mathrm{d}W_t, \qquad 0 \le t \le T \tag{3.3.37}$$

with initial condition $X_0 = 1$. The solution of this is given by $X_t = e^{W_t - \frac{t}{2}}$, the stochastic (or Doléans-Dade) exponential *of the Brownian motion. To show this, consider any C^2 function $\varphi(x)$; Itô's rule then gives*

$$\begin{aligned}\mathrm{d}\varphi(X_t) &= \varphi'(X_t)\,\mathrm{d}X_t + \frac{1}{2}\varphi''(X_t)(\mathrm{d}X_t)^2 &(3.3.38)\\ &= \frac{1}{2}\varphi''(X_t)X_t^2\,\mathrm{d}t + \varphi'(X_t)X_t\,\mathrm{d}W_t. &(3.3.39)\end{aligned}$$

In particular, taking $\varphi(x) = \log x$, we get

$$\mathrm{d}\log X_t = -\frac{1}{2}\,\mathrm{d}t + \mathrm{d}W_t, \tag{3.3.40}$$

which is readily integrated to give $\log X_t = W_t - \frac{t}{2}$. The claim then follows.

Example 7 (Geometric Brownian motion)**.** *Adding a linear drift to* (3.3.37) *and rescaling the Brownian motion gives*

$$\mathrm{d}X_t = \mu X_t \,\mathrm{d}t + \sigma X_t \,\mathrm{d}W_t, \tag{3.3.41}$$

where $\mu \in \mathbb{R}$ *and* $\sigma > 0$ *are known as the* drift *and the* volatility *parameters. The solution of* (3.3.41) *with the initial condition* $X_0 = 0$ *is known as the* geometric Brownian motion*. To obtain an explicit formula, we once again consider* $\log X_t$*:*

$$\mathrm{d}\log X_t = \frac{1}{X_t}\,\mathrm{d}X_t - \frac{1}{2X_t^2}(\mathrm{d}X_t)^2 \tag{3.3.42}$$

$$= \left(\mu - \frac{\sigma^2}{2}\right)\mathrm{d}t + \sigma\,\mathrm{d}W_t, \tag{3.3.43}$$

which gives

$$X_t = \exp\left(\sigma W_t + \left(\mu - \frac{\sigma^2}{2}\right)t\right). \tag{3.3.44}$$

Geometric Brownian motion is used in mathematical finance to model stock prices in the so-called Black–Scholes model.

3.3.2 Coordinate changes and the Stratonovich integral

When working with dynamical systems modeled by ordinary differential equations, the choice of the coordinate system often makes a difference when it comes to ease of analysis. Consider an ODE

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = f(x(t)) \tag{3.3.45}$$

describing the evolution of an n-dimensional state vector $x(t)$, and let a smooth (e.g., C^1) function $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ be given. Then the evolution of $y(t) := \varphi(x(t))$ is easily described using the chain rule:

$$\frac{\mathrm{d}}{\mathrm{d}t}y(t) = \frac{\mathrm{d}}{\mathrm{d}t}\varphi(x(t)) \tag{3.3.46}$$

$$= \frac{\partial\varphi}{\partial x}(x(t))\frac{\mathrm{d}}{\mathrm{d}t}x(t) \tag{3.3.47}$$

$$= \frac{\partial\varphi}{\partial x}(x(t))f(x(t)), \tag{3.3.48}$$

where $\frac{\partial\varphi}{\partial x}$ is the Jacobian of φ, i.e., the $n \times n$ matrix of all its first-order partial derivatives. If φ is invertible, then we obtain an equivalent model in terms of $y(t)$ only:

$$\frac{\mathrm{d}}{\mathrm{d}t}y(t) = \tilde{f}(y(t)), \tag{3.3.49}$$

where $\tilde{f}(y) := \frac{\partial\varphi}{\partial x}(\varphi^{-1}(y))f(\varphi^{-1}(y))$. This procedure is useful, for example, when passing to the new coordinates exposes some favorable properties of the dynamics or allows one to obtain an explicit solution.

At this point it should not be surprising that the situation would be different when working with SDEs. Indeed, if we start with an n-dimensional SDE

$$\mathrm{d}X_t = f(X_t)\,\mathrm{d}t + g(X_t)\,\mathrm{d}W_t \tag{3.3.50}$$

driven by an m-dimensional Brownian motion and consider the process $Y_t = \varphi(X_t)$, then Itô's rule gives

$$\mathrm{d}Y_t^i = \frac{\partial \varphi^i}{\partial x}(X_t)\,\mathrm{d}X_t + \frac{1}{2}(\mathrm{d}X_t)^T \frac{\partial^2 \varphi^i}{\partial x^2}(X_t)\,\mathrm{d}X_t, \qquad i = 1, \dots, n \tag{3.3.51}$$

where $\frac{\partial \varphi^i}{\partial x} = \left(\frac{\partial \varphi^i}{\partial x^1}, \dots, \frac{\partial \varphi^i}{\partial x^n}\right)$ is the row vector of the first-order partial derivatives of the ith coordinate of $y = \varphi(x)$, while $\frac{\partial^2 \varphi^i}{\partial x^2} = \left(\frac{\partial^2 \varphi^i}{\partial x^j \partial x^k}\right)_{1 \le j,k \le n}$ is the $n \times n$ matrix of the second-order partial derivatives of φ^i. As before, only the first term, involving $\frac{\partial \varphi^i}{\partial x}$, matches the chain rule of ordinary calculus, but we also have the other term involving second-order derivatives of φ.

Now, it is possible to define stochastic integrals in a manner that would preserve the rules of ordinary calculus. This was done independently by Stratonovich and, independently, by Fisk. A nice way to introduce the Stratonovich integral, different from the original construction, is to see what needs to be modified in Itô's definition in order to get back the rules of ordinary calculus. (We will also see that this convenience comes at a price.) To keep things simple, we will consider the one-dimensional case. Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a C^1 function; then Itô's rule gives

$$\int_0^t \varphi(W_s)\,\mathrm{d}W_s = \Phi(W_t) - \Phi(W_0) - \frac{1}{2}\int_0^t \varphi'(W_s)\,\mathrm{d}s, \tag{3.3.52}$$

where $\Phi(x) := \int_0^x \varphi(y)\,\mathrm{d}y$ and the integral on the left-hand side is the Itô integral. Now, if we *define* another type of stochastic integral, namely

$$\int_0^t \varphi(W_s) \circ \mathrm{d}W_s := \int_0^t \varphi(W_s)\,\mathrm{d}W_s + \frac{1}{2}\int_0^t \varphi'(W_s)\,\mathrm{d}s, \tag{3.3.53}$$

then the fundamental theorem of ordinary calculus holds by construction:

$$\int_0^t \varphi(W_s) \circ \mathrm{d}W_s = \Phi(W_t) - \Phi(W_0). \tag{3.3.54}$$

This procedure may look somewhat contrived, but we can generalize it using Itô's product rule. To that end, let X_t and $\tilde{X}_t$ be two Itô processes. Then the product rule tells us that

$$X_t\tilde{X}_t - X_0\tilde{X}_0 = \int_0^t X_s\,\mathrm{d}\tilde{X}_s + \int_0^t \tilde{X}_s\,\mathrm{d}X_s + \int_0^t \mathrm{d}[X, \tilde{X}]_s, \tag{3.3.55}$$

where $[X, \tilde{X}]_t$ is the joint quadratic variation of X_t and $\tilde{X}_t$. If we split the joint quadratic variation term equally between the first two Itô integrals on the right-hand side, then we can define

$$\int_0^t X_s \circ \mathrm{d}\tilde{X}_s := \int_0^t X_s\,\mathrm{d}\tilde{X}_s + \frac{1}{2}\int_0^t \mathrm{d}[X, \tilde{X}]_s \tag{3.3.56}$$

which gives us something that looks like the product rule of ordinary calculus:

$$\int_0^t X_s \circ \mathrm{d}\tilde{X}_s + \int_0^t \tilde{X}_s \circ \mathrm{d}X_s = X_t\tilde{X}_t - X_0\tilde{X}_0. \tag{3.3.57}$$

We can then take (3.3.56) as the definition of the *Stratonovich integral* of X w.r.t. $\tilde{X}$, and it easy to verify that (3.3.53) is indeed a special case of (3.3.56):

$$\begin{aligned} \mathrm{d}[\varphi(W), W]_t &= \mathrm{d}\varphi(W_t)\,\mathrm{d}W_t && (3.3.58)\\ &= (\varphi'(W_t)\,\mathrm{d}W_t + \frac{1}{2}\varphi''(W_t)\,\mathrm{d}t)\,\mathrm{d}W_t && (3.3.59)\\ &= \varphi'(W_t)\,\mathrm{d}t. && (3.3.60) \end{aligned}$$

We can extend the above definition to C^1 functions of Itô processes; e.g., if h is such a function, then we set

$$\int_0^t h(X_s) \circ \mathrm{d}\tilde{X}_s := \int_0^t h(X_s)\,\mathrm{d}\tilde{X}_s + \frac{1}{2}\int_0^t h'(X_s)\,\mathrm{d}[X, \tilde{X}]_s \tag{3.3.61}$$

(this definition coincides with (3.3.56) if $h(X_t)$ happens to be an Itô process). Thus, in particular, if φ is C^2 and if we take $\tilde{X} = X$, then

$$\begin{aligned} \int_0^t \varphi'(X_s) \circ \mathrm{d}X_s &= \int_0^t \varphi'(X_s)\,\mathrm{d}X_s + \frac{1}{2}\int_0^t \varphi''(X_s)\,\mathrm{d}[X, X]_s && (3.3.62)\\ &= \varphi(X_t) - \varphi(X_0), && (3.3.63) \end{aligned}$$

since, by Itô's rule, the quantity on the right-hand side is the integral from 0 to t of the total Itô differential $\mathrm{d}\varphi(X_s)$.

Remark 4. *The original definition of Fisk and Stratonovich for the integral $\int_0^t h(X_s) \circ \mathrm{d}W_s$ was as the limit in probability of sums of the form*

$$\sum_{i=0}^{n-1} h\Big(\frac{1}{2}(X_{t_{i+1}} + X_{t_i})\Big)(W_{t_{i+1}} - W_{t_i}) \tag{3.3.64}$$

for any sequence of partitions of $[0, t]$ that get increasingly finer as $n \to \infty$. In particular, if $X_t = W_t$ and $h(x) = x$, we have

$$\frac{1}{2}(W_{t_{i+1}} + W_{t_i})(W_{t_{i+1}} - W_{t_i}) = \frac{1}{2}(W_{t_{i+1}}^2 - W_{t_i}^2), \tag{3.3.65}$$

so the sum in (3.3.64) telescopes to $\frac{1}{2}W_t^2 - \frac{1}{2}W_0^2$.

We can now write down Stratonovich SDEs. For example, let X_t be a process described by a scalar Itô SDE

$$\mathrm{d}X_t = f(X_t)\,\mathrm{d}t + g(X_t)\,\mathrm{d}W_t, \tag{3.3.66}$$

where f and g satisfy the regularity conditions of Theorem 9, and g is also C^1. Then it can also be described by the Stratonovich SDE

$$\mathrm{d}X_t = \hat{f}(X_t)\,\mathrm{d}t + g(X_t) \circ \mathrm{d}W_t, \tag{3.3.67}$$

where the drift $\hat{f}$ is related to f and g through

$$\hat{f}(x) = f(x) - \frac{1}{2} g(x) g'(x) \tag{3.3.68}$$

(the extra term subtracted from f is usually called the *Itô correction*). This is easy to show: integrating (3.3.66) and using (3.3.61), we have

$$\begin{aligned} X_t &= X_0 + \int_0^t f(X_s)\,\mathrm{d}s + \int_0^t g(X_s)\,\mathrm{d}W_s && (3.3.69)\\ &= X_0 + \int_0^t f(X_s)\,\mathrm{d}s + \int_0^t g(X_s) \circ \mathrm{d}W_s - \frac{1}{2}\int_0^t g'(X_s)\,\mathrm{d}[X,W]_s && (3.3.70)\\ &= X_0 + \int_0^t \left(f(X_s) - \frac{1}{2} g(X_s) g'(X_s) \right) \mathrm{d}s + \int_0^t g(X_s) \circ \mathrm{d}W_s && (3.3.71)\\ &= X_0 + \int_0^t \hat{f}(X_s)\,\mathrm{d}s + \int_0^t g(X_s) \circ \mathrm{d}W_s. && (3.3.72) \end{aligned}$$

The main benefit of working with Stratonovich SDEs is that they respect the chain rule of ordinary calculus. Indeed, if φ is a C^2 function, then $Y_t = \varphi(X_t)$ satisfies

$$\begin{aligned} \mathrm{d}Y_t &= \left(\varphi'(X_t) f(X_t) + \frac{1}{2}\varphi''(X_t) g^2(X_t) \right) \mathrm{d}t + \varphi'(X_t) g(X_t)\,\mathrm{d}W_t && (3.3.73)\\ &= \left(\varphi'(X_t) f(X_t) + \frac{1}{2}\varphi''(X_t) g^2(X_t) \right) \mathrm{d}t + \varphi'(X_t) g(X_t) \circ \mathrm{d}W_t - \frac{1}{2}(\varphi'(X_t) g(X_t))' g(X_t)\,\mathrm{d}t && (3.3.74)\\ &= \varphi'(X_t) \left(\hat{f}(X_t)\,\mathrm{d}t + g(X_t) \circ \mathrm{d}W_t \right), && (3.3.75) \end{aligned}$$

where we have first used Itô's rule, then converted Itô differentials to Stratonovich differentials, and then collected terms to get the final expression. The main downside is that, unlike Itô integrals, Stratonovich integrals are not martingales.

3.4 SDEs and models of engineering systems with random disturbances

Our digression on the stochastic calculus of Stratonovich was motivated by the failure of the familiar change-of-variables formula when working with Itô differentials. On the one hand, this should not come as a surprise: One cannot interpret Itô SDEs as differential equations in the usual sense, not the least because their solutions (being diffusion processes) do not have differentiable sample paths, unlike the solutions of ODEs where differentiability is expected more or less by design. On the other hand, the way we have arrived at Itô SDEs was via Doob's solution of the realization problem for diffusion processes specified through their transition probability functions. Thus, as far as the questions of describing the temporal evolution of means, covariances, and other expected values are concerned, the stochastic calculus of Itô is a convenient framework, especially given various useful properties of the Itô integral.

On the other hand, again by appealing to the results of Doob and Itô, our internalist models of diffusion processes make fundamental use of Brownian motion, which is a mathematical idealization

of various physical noise processes. Thus, one should think of diffusion processes as themselves idealizations of the state trajectories of physical or engineering systems subject to random excitations or disturbances. This was the original motivation of the early phenomenological work of Pontryagin, Andronov, and Vitt, and it was the motivation of the work of Stratonovich as well. In particular, realistic models of stochastic systems involve disturbance processes whose correlation times are much shorter than the characteristic time constants of the signals generated by the system. In such instances, we may hope that the signals can be described by solutions of ODEs of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}X_t = f(X_t) + g(X_t)\xi_t, \tag{3.4.1}$$

where ξ_t is a "physical" (e.g., wideband) noise process. If this process is sufficiently well-behaved, then it should be possible to obtain a solution of (3.4.1) using ordinary calculus. Such a solution will be (at least once) differentiable as a function of time, which is consistent with empirical observations of many types of physical and engineering systems. Of course, the catch is that the corresponding process will generally not be Markov due to the disturbance having nonzero memory. Thus, it is reasonable to consider a *family* of disturbance processes parametrized by something like a "maximum frequency" $\alpha > 0$ such that the processes become more and more "white-noise-like" as $\alpha \to \infty$, and then ask whether the solutions X_t^α of the corresponding family of ODEs

$$\frac{\mathrm{d}}{\mathrm{d}t}X_t^\alpha = f(X_t^\alpha) + g(X_t^\alpha)\xi_t^\alpha \tag{3.4.2}$$

converge in some sense to a diffusion process, i.e., a solution of an SDE. We would also like to know the drift and the diffusion matrix of this process and in what sense (Itô or Stratonovich) we should interpret this SDE.

3.4.1 The Wong–Zakai theorem

We start by giving one such result, which is relatively easy to state and to prove (at least in the scalar setting). Thus, consider the family of ODEs (3.4.2) on the finite time interval $[0, T]$, all having a common initial condition X_0. We assume that, for each value of α, the disturbance process $(\xi_t^\alpha)_{t\in[0,T]}$ is sufficiently regular to guarantee that (3.4.2) has a unique solution in the usual sense. We also assume that, as $\alpha \to \infty$, the disturbance processes (ξ_t^α) become better and better approximations to white noise in the sense that there exists a Brownian motion $(W_t)_{t\in[0,T]}$, such that

$$\sup_{t\in[0,T]} |W_t^\alpha - W_t| \xrightarrow{\alpha\to\infty} 0 \text{ a.s.}, \qquad W_t^\alpha := \int_0^t \xi_s^\alpha \,\mathrm{d}s. \tag{3.4.3}$$

Finally, we make the following assumptons on the functions f and g:

1. f and g are both Lipschitz continuous and bounded.

2. g is C^1, and the function $x \mapsto g(x)g'(x)$ is Lipschitz continuous (the prime denotes differentiation w.r.t. x);

3. there exists a constant $c > 0$, such that $g(x) \geq c$ for all x.

Theorem 10 (Wong–Zakai)**.** *Under the above assumptions, the solutions X_t^α of the ODEs (3.4.2) converge, as $\alpha \to \infty$, a.s. to the unique solution X_t of the Stratonovich SDE*

$$\mathrm{d}X_t = f(X_t) + g(X_t) \circ \mathrm{d}W_t \tag{3.4.4}$$

with the same initial condition, i.e.,

$$\sup_{t\in[0,T]} |X_t^\alpha - X_t| = 0 \ \text{a.s.} \tag{3.4.5}$$

Remark 5. *It is important to keep in mind that, when we express (3.4.4) in the Itô form, the drift willl pick up the Itô correction term:*

$$\mathrm{d}X_t = \left(f(X_t) + \frac{1}{2} g(X_t) g'(X_t) \right) \mathrm{d}t + g(X_t)\, \mathrm{d}W_t. \tag{3.4.6}$$

When g is constant, i.e., $g(x) = \sigma$ for some $\sigma > 0$, the Stratonovich and the Itô SDEs coincide.

Proof. First, it is easy to verify that, under our assumptions on f and g, the SDE (3.4.4) has a unique solution X_t. Define now the function

$$\varphi(x) := \int_0^x \frac{1}{g(z)}\, \mathrm{d}z, \tag{3.4.7}$$

so that $\varphi'(x) = \frac{1}{g(x)}$ and $\varphi''(x) = -\frac{g'(x)}{g^2(x)}$. Then, for $Y_t^\alpha := \varphi(X_t^\alpha)$ and $Y_t := \varphi(X_t)$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t} Y_t^\alpha = \frac{f(X_t^\alpha)}{g(X_t^\alpha)} + \xi_t^\alpha \tag{3.4.8}$$

and

$$\mathrm{d}Y_t = \frac{f(X_t)}{g(X_t)}\, \mathrm{d}t + \mathrm{d}W_t, \tag{3.4.9}$$

where (3.4.8) follows form the ordinary chain rule applied to the solution of the ODE (3.4.2), while (3.4.9) follows from Itô's rule. Integrating (3.4.8) and (3.4.9) and using the fact that $X_0^\alpha = X_0$ for all α, we get

$$Y_t^\alpha - Y_t = \int_0^t \left(\frac{f(X_t^\alpha)}{g(X_t^\alpha)} - \frac{f(X_t)}{g(X_t)} \right) \mathrm{d}t + W_t^\alpha - W_t. \tag{3.4.10}$$

From our assumptions on f and g it easily follows that

$$|\varphi(x) - \varphi(z)| = \left| \int_z^x \frac{1}{g(v)}\, \mathrm{d}v \right| \geq \frac{1}{C_1} |x - z| \tag{3.4.11}$$

and

$$\left| \frac{f(x)}{g(x)} - \frac{f(z)}{g(z)} \right| \leq \frac{|f(x) - f(z)|}{g(x)} + |f(z)| \frac{|g(x) - g(z)|}{g(x) g(z)} \leq C_2 |x - z| \tag{3.4.12}$$

for some $C_1, C_2 > 0$. Using these inequalities together with (3.4.10), we obtain

$$|X_t^\alpha - X_t| \leq C_1 C_2 \int_0^t |X_s^\alpha - X_s|\, \mathrm{d}s + C_1 \sup_{0\leq t\leq T} |W_t^\alpha - W_t|, \tag{3.4.13}$$

whence, by Gronwall's lemma, it follows that

$$\sup_{0\leq t\leq T} |X_t^\alpha - X_t| \leq C_1 e^{C_1 C_2 T} \sup_{0\leq t\leq T} |W_t^\alpha - W_t|. \tag{3.4.14}$$

Taking the limit of both sides as $\alpha \to \infty$, we obtain the claimed result. □

3.4.2 Other models of physical noise processes

The theorem of Wong and Zakai is only a particular result among many more general results of this nature. For example, one can consider a wide class of noise processes ξ_t in (3.4.1) that are obtained from Brownian motion by linear filtering, i.e., there exists a vector Brownian motion (W_t) and a deterministic family of matrices $h(t,s)$ of appropriate shape, such that

$$\xi_t = \int_0^t h(t,s)\,\mathrm{d}W_s. \tag{3.4.15}$$

Each such process is evidently Gaussian, and many types of random phenomena in engineering systems can be modeled in this way. Here are a few examples:

1. *Approximately white stationary noise.* Let ξ_t be a scalar Gaussian stationary process whose power spectral density $S(\omega)$ satisfies the so-called *Paley–Wiener condition*

$$\int_{-\infty}^{\infty} \frac{\log S(\omega)}{1+\omega^2}\,\mathrm{d}\omega < \infty. \tag{3.4.16}$$

   Then there exist a scalar Brownian motion $(W_t)_{-\infty<t<\infty}$ and a deterministic function $h(t)$, such that ξ_t can be represented as

$$\xi_t = \int_{-\infty}^{t} h(t-s)\,\mathrm{d}W_s. \tag{3.4.17}$$

   In the above representation, ξ_t depends on the entire past $(W_s)_{-\infty<s\le t}$ of the Brownian motion. However, if ξ_t has very short correlation time, then $h(t-s)$ will decay rapidly as $t-s\to\infty$, and in that case we can approximate ξ_t by cutting off the integration in (3.4.17) at $s=0$.

2. *The Ornstein–Uhlenbeck process.* The diffusion process that solves the linear Itô SDE

$$\mathrm{d}\xi_t = F\xi_t\,\mathrm{d}t + G\,\mathrm{d}W_t, \qquad \xi_0 = 0 \tag{3.4.18}$$

   has the explicit representation

$$\xi_t = \int_0^t e^{(t-s)F} G\,\mathrm{d}W_s, \tag{3.4.19}$$

   as shown in Example 5.

3. *Piecewise constant processes.* Consider the process ξ_t which is constant over intervals of length $\delta>0$, and the values that it takes in nonoverlapping intervals are independent zero-mean Gaussian random vectors. Such a process can be represented in the form (3.4.15) with

$$h(t,s) = \begin{cases} \frac{1}{\delta} I, & (n-1)\delta \le s \le n\delta \le t < (n+1)\delta,\ n=1,2,\dots \\ 0, & \text{otherwise} \end{cases}. \tag{3.4.20}$$

One can construct other examples of this type. The bottom line is that many types of noise processes ξ_t can be represented in the form (3.4.15), where the matrices $h(t,s)$ are such that ξ_t has a.s. piecewise continuous sample paths.

Now, just as we had done to prove the Wong–Zakai theorem, we can consider a family (ξ_t^α) of such processes, where $\alpha>0$ is some sort of an "effective maximum frequency" parameter, and we assume that there exist constant matrices A, B and a Brownian motion (W_t), such that:

- for all $0 < t \le T$,

$$\lim_{\alpha\to\infty} \frac{1}{t} \int_0^t \int_0^s \mathbf{E}[\xi_s^\alpha (\xi_r^\alpha)^T] \, \mathrm{d}r \, \mathrm{d}s = A \tag{3.4.21}$$

- for all $0 < t \le T$,

$$\lim_{\alpha\to\infty} \mathbf{E} \left| \int_0^t \xi_s \, \mathrm{d}s - BW_t \right|^2 = 0 \tag{3.4.22}$$

  (the matrices A and B are then related by $BB^T = A + A^T$).

Then, under some additional technical conditions, it can be shown that the solutions X_t^α of (3.4.2) for $0 \le t \le T$ with zero initial condition $X_0^\alpha = 0$ converge as $\alpha \to \infty$ to the solution of the Stratonovich SDE

$$\mathrm{d}X_t = \hat{f}(X_t) \, \mathrm{d}t + \hat{g}(X_t) \circ \mathrm{d}W_t, \qquad X_0 = 0 \tag{3.4.23}$$

with

$$\hat{f}^i(x) = f^i(x) + \frac{1}{2} \sum_{j,k,\ell} g^{jk}(x) \frac{\partial}{\partial x^j} g^{ik}(x) (A_{k\ell} - A_{\ell k}) \tag{3.4.24}$$

and

$$\hat{g}^{ij}(x) = g^{ij}(x) B^{ij}. \tag{3.4.25}$$

The convergence is in mean-square sense, such that $\mathbf{E}|X_t^\alpha - X_t|^2 = O(\alpha^{-1})$.

3.5 Problems

1. Let $(W_t)_{t\ge0}$ be a standard Brownian motion adapted to a filtration $(\mathcal{F}_t)_{t\ge0}$. Prove that $M_t = \exp(\mu W_t - \frac{\mu^2}{2} t)$ is an $\mathcal{F}_t$-martingale for any value of μ.

2. Give a direct proof that $X_t = \exp(W_t - \frac{t}{2})$ is the unique solution of the Itô SDE $\mathrm{d}X_t = X_t \, \mathrm{d}W_t$ with the initial condition $X_0 = 1$.
Hint: Use Itô's product rule.

3. Consider the Ornstein–Uhlenbeck SDE $\mathrm{d}X_t = FX_t \, \mathrm{d}t + G \, \mathrm{d}W_t$, as in Example 5. For an initial condition X_0 independent of the Brownian motion W_t, its solution is given by

$$X_t = e^{tF} X_0 + \int_0^t e^{(t-s)F} G \, \mathrm{d}W_s. \tag{3.5.1}$$

(i) Suppose that X_0 is a Gaussian random vector. Prove that X_t is a Gaussian process and find its mean $\mu(t) := \mathbf{E}[X_t]$ and covariance matrix $K(s,t) := \mathbf{E}[(X_s - \mu(s))(X_t - \mu(t))^T]$.

(ii) Suppose that F is a *Hurwitz matrix*, i.e., all of its eigenvalues have negative real parts. Find the asymptotic mean $\bar{\mu} := \lim_{t\to\infty} \mu(t)$ and the asymptotic covariance matrix $\bar{K} := \lim_{t\to\infty} K(t,t)$.

4. Solve the one-dimensional Itô SDE

$$\mathrm{d}X_t = (X_t^3 - X_t)\,\mathrm{d}t + (1 - X_t^2)\,\mathrm{d}W_t \tag{3.5.2}$$

with the initial condition $X_0 = 0$.

Hint: Try a solution of the form $X_t = \varphi(W_t)$ for a C^2 function φ and use Itô's rule. Problem 6 in Section 2.4 may be useful as well.

5. Use Itô calculus to prove Theorem 2.

6. According to the Nyquist–Johnson model of thermal noise, the current I_t and the voltage V_t in a resistor with conductance[1] g are related via the formula

$$I_t = gV_t + \sqrt{2kgT}\xi_t, \tag{3.5.3}$$

where ξ_t is a white noise process (the formal derivative of the standard one-dimensional Brownian motion), k is Boltzmann's constant, and T is the absolute temperature. Thus, a Nyquist–Johnson resistor is modeled by an ideal (noiseless) resistor in series with a white-noise voltage source (or in parallel with a white-noise current source). A circuit consisting of a Nyquist–Johnson resistor in series with a linear capacitor with (possibly time-varying) capacitance $c(t)$ can be described by an Itô SDE

$$\mathrm{d}(c(t)V_t) = -gV_t\,\mathrm{d}t + \sqrt{2kgT}\,\mathrm{d}W_t, \qquad t \geq 0 \tag{3.5.4}$$

where V_t is the voltage across the capacitor, and we assumme that $\mathbf{E}[V_0^2] < \infty$.

(i) Consider first the case when the capacitance is constant, i.e., $c(t) \equiv c > 0$ for all t, and find the steady-state energy stored in the capacitor,

$$\bar{e} := \lim_{t\to\infty} \frac{1}{2}\mathbf{E}[cV_t^2]. \tag{3.5.5}$$

(ii) Now suppose that the capacitor is time-varying, and its capacitance $c(t) > 0$ is evolving deterministically according to the ODE $\frac{\mathrm{d}}{\mathrm{d}t}c(t) = u(t)$, where $u(t)$ is a given function of time (we can think of it as a control input). Write down an Itô SDE for the voltage V_t across the capacitor.

(iii) Based on the result of part (ii), write down an ODE for $e(t) := \frac{1}{2}\mathbf{E}[c(t)V_t^2]$, the average energy stored in the capacitor at time t, and solve it for $e(t)$.

Hint: You should end up with an ODE of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}e(t) = -a(t)e(t) + b(t),$$

where $a(\cdot)$ and $b(\cdot)$ are some fixed functions of t; you can then use the method of integrating factors.

(iv) Consider now the case when $c(0) = c > 0$ and $u(t) = \lambda$ for some $\lambda > 0$. Find an explicit expression for $e(t)$ from part (iii) and for its steady-state value $\bar{e} := \lim_{t\to\infty} e(t)$. What happens to $\bar{e}$ as $\lambda \to 0$?

[1] Defined as the reciprocal of resistance.

3.6 Notes and further reading

A detailed exposition of the system modeling philosophy involving externalist and internalist descriptions can be found in [Wil89], and [Pic91] provides a discussion of stochastic realization theory. The derivation of the martingale representation of a diffusion process in Section 3.1 is based on Chapter 11 of [Nel67], which itself was based on the book of Doob [Doo53]. The (largely informal) presentation of martingale theory and related concepts was inspired by the survey article of Wong [Won73].

Interesting comparative discussions of Itô vs. Stratonovich integrals can be found in [Mor69] (from an engineering perspective) and [Kam81] (from a physics perspective).

The Wong–Zakai theorem has its origins in [WZ65a, WZ65b]. The discussion of physical noise processes obtained by linear filtering of Brownian motion is loosely based on [Cla66].

Chapter 4

Stochastic calculus in path space

So far, our probabilistic analysis of diffusion processes was largely local: Given a diffusion process (X_t), we were able to write down expressions for expectations $\mathbf{E}[\varphi(X_t)|X_s = x]$ in terms of the drift and the diffusion coefficient of the process. However, early on we were emphasizing the fact that even the most basic diffusion process—the Brownian motion—has continuous sample paths. This is a *global* property of the entire path (although continuity is defined locally, in a neighborhood of every time t). In other words, when we talk about continuity of sample paths, we are viewing the process as an element of a function space. For instance, if $(W_t)_{t\in[0,T]}$ is a standard d-dimensional Brownian motion, then its path is a random element of the space $C([0,T];\mathbb{R}^d)$ of continuous functions from $[0,T]$ into $\mathbb{R}^d$, and we can talk about expectations of real-valued functionals of the entire path, for example

$$\mathbf{E}\left[\sup_{t\in[0,T]}|W_t|\right] \quad \text{or} \quad \mathbf{E}\left[\int_0^T \mathbf{1}_{\{W_t\geq 0\}}\,\mathrm{d}t\right].$$

In order to do this properly, we need to specify an appropriate sample space Ω and a σ-algebra of events on it. We have already selected Ω as $C([0,T];\mathbb{R}^d)$, and to define the σ-algebra we will make use of the fact that $\Omega = C([0,T];\mathbb{R}^d)$ comes equipped with a distance: If $\omega = (\omega_t)_{t\in[0,T]}$ and $\tilde{\omega} = (\tilde{\omega}_t)_{t\in[0,T]}$ are two continuous functions of $t \in [0,T]$ with values in $\mathbb{R}^d$, then we take

$$d(\omega,\tilde{\omega}) := \sup_{t\in[0,T]}\max_{1\leq i\leq d}|\omega_t^i - \tilde{\omega}_t^i|,$$

where ω_t^i is the ith coordinate of $\omega_t \in \mathbb{R}^d$. We then let $\mathcal{B}$ denote the smallest σ-algebra containing all sets of the form $\{\omega\in\Omega : d(\omega,\nu) < r\}$ for all $\nu\in\Omega$ and all $r > 0$. This is called the *Borel σ-algebra*, in analogy with the Borel σ-algebra on $\mathbb{R}^d$, which is the smallest σ-algebra containing all open balls in $\mathbb{R}^d$. There is also an alternative characterization of $\mathcal{B}$: It is the smallest σ-algebra on Ω containing all sets of the form

$$C = \left\{\omega\in\Omega : \omega_{t_i}\in A_i,\, i = 1,\ldots,n\right\}, \tag{4.0.1}$$

for all n, all choices of times $0 < t_1 < \ldots < t_n \leq T$, and all Borel sets $A_i \subseteq \mathbb{R}^d$. Sets of this form are referred to as the *cylinder sets*, and they usually have nice operational interpretations: Suppose, for example, that $\omega_t = (\omega_t^1,\ldots,\omega_t^d)$ is a collection of voltage waveforms measured at d nodes of

an electric circuit. Then the set C in (4.0.1) describes the situation when the vectors of voltages measured at times $t_1, \dots, t_n$ fall within some given sets $A_1, \dots, A_n \subset \mathbb{R}^d$.

With these definitions, a d-dimensional $\mathbb{R}^d$-valued random process $(X_t)_{t\in[0,T]}$ with continuous sample paths defined on an arbitrary probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\boldsymbol{P}})$ can be viewed as a random element of Ω. To do this, we first define a mapping $\boldsymbol{X} : \tilde{\Omega} \to C([0,T];\mathbb{R}^d)$, such that $\omega := \boldsymbol{X}(\tilde{\omega})$ maps each $t \in [0,T]$ to $X_t(\tilde{\omega})$. This then *induces* a probability measure $\boldsymbol{P}$ on $(\Omega, \mathcal{B})$ by $\boldsymbol{P}(A) := \tilde{\boldsymbol{P}}(\boldsymbol{X}^{-1}(A))$ for all $A \in \mathcal{B}$. Alternatively, we can introduce probability measures directly on $(\Omega, \mathcal{B})$. In particular, we can let $\boldsymbol{P}$ denote the probability measure on $(\Omega, \mathcal{B})$ that corresponds to the standard d-dimensional Brownian motion; thus, for the set C defined in (4.0.1), we will have

$$\boldsymbol{P}(C) = \int_{A_1\times\cdots\times A_n} \prod_{i=1}^{n} \frac{1}{(2\pi(t_i - t_{i-1})^{d/2}} \exp\left(-\frac{|x_i - x_{i-1}|^2}{2(t_i - t_{i-1})}\right) \mathrm{d}x_1 \dots \mathrm{d}x_n, \tag{4.0.2}$$

where we have taken $(x_0, t_0) = (0, 0)$. It can also be shown that any measurable function $f : \Omega \to \mathbb{R}$ (i.e., f is such that all the sets of the form $\{\omega \in \Omega : f(\omega) < a\}$ for $a \in \mathbb{R}$ belong to $\mathcal{B}$) can be approximated arbitrarily closely by sums of indicator functions of cylinder sets of the form (4.0.1). We will also say things like "$(W_t)_{t\in[0,T]}$ is a $\boldsymbol{P}$-Brownian motion" to indicate this, with the understanding that changing the probability measure from $\boldsymbol{P}$ to some other $\boldsymbol{Q}$ will also change the probability law of the process.

4.1 Cameron–Martin–Girsanov theory

With these preliminaries out of the way, let us consider the following problem (let's consider the scalar case to keep things simple): We have a diffusion process $(X_t)_{t\in[0,T]}$, which, as a random element of $\Omega = C([0,T])$, has probability law $\boldsymbol{Q}$. Suppose that $\boldsymbol{Q}$ is *absolutely continuous* w.r.t. $\boldsymbol{P}$, the probability law of the Brownian motion. This means that, for any $A \in \mathcal{B}$ such that $\boldsymbol{P}(A) = 0$, we also have $\boldsymbol{Q}(A) = 0$. A good example of such a process would be an Itô process of the form

$$X_t = \int_0^t F_s \, \mathrm{d}s + W_t, \tag{4.1.1}$$

where (W_t) is a $\boldsymbol{P}$-Brownian motion and (F_t) is an adapted process (and, in fact, this is the only possibility). Then, by absolute continuity, there exists some nonnegative function $g : \Omega \to \mathbb{R}_+$ (called the *Radon–Nikodym derivative* of $\boldsymbol{Q}$ w.r.t. $\boldsymbol{P}$ and often denoted by $\frac{\mathrm{d}\boldsymbol{Q}}{\mathrm{d}\boldsymbol{P}}$) such that

$$\mathbf{E}_{\boldsymbol{Q}}[\mathbf{1}_A(\boldsymbol{X})] = \mathbf{E}_{\boldsymbol{P}}[\mathbf{1}_A(\boldsymbol{W})g(\boldsymbol{W})], \qquad A \in \mathcal{B}. \tag{4.1.2}$$

A representation like this has obvious benefits, one being that the computation of expectations w.r.t. $\boldsymbol{P}$ is (presumably) easier than working with $\boldsymbol{Q}$. Of course, this hinges on the availability of an explicit expression for g. This is precisely the content of so-called *Cameron–Martin–Girsanov theory* (or just Girsanov theory), although its reach goes far beyond Itô processes.

4.1.1 Motivation: importance sampling

Let's consider, by way of motivation, a very simple one-dimensional example. Let W be a standard Gaussian random variable, i.e., $W \sim \mathcal{N}(0,1)$ and let $X = \mu + W$ for some $\mu \in \mathbb{R}$. Then, for any

well-behaved (say, bounded) function f, we have

$$\mathbf{E}[f(W)] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-\frac{1}{2}x^2} \, \mathrm{d}x. \tag{4.1.3}$$

Expanding the square $x^2 = ((x-\mu)+\mu)^2$ and rearranging, we can express this in another way:

$$\mathbf{E}[f(W)] = \frac{1}{\sqrt{2\pi}} f(x) e^{\mu x - \frac{1}{2}\mu^2} e^{-\frac{1}{2}(x-\mu)^2} \, \mathrm{d}x \tag{4.1.4}$$

$$= \mathbf{E}[f(X) e^{\mu X - \frac{1}{2}\mu^2}], \tag{4.1.5}$$

so we can "shift the mean" by a simple exponential transformation. Stated in this way, this idea borders on triviality, but it has practical consequences. For example, consider the function $f(x) = \mathbf{1}_{\{x \geq 20\}}$. Then

$$\mathbf{E}[f(W)] = \mathbf{P}[W \geq 20] \leq e^{-(20)^2/2} \approx 1.4 \times 10^{-87}. \tag{4.1.6}$$

Suppose we wanted to *simulate* the event $\{W \geq 20\}$—the usual way is to generate a large number n of i.i.d. samples from $\mathcal{N}(0,1)$ and to look at the fraction of times the threshold 20 was exceeded:

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{W_i \geq 20\}} \equiv \frac{1}{n} \sum_{i=1}^{n} f(W_i). \tag{4.1.7}$$

Of course we know that, by the law of large numbers, this fraction converges to its expected value $\mathbf{P}[W \geq 20]$, but a simple calculation shows that we would have to generate a huge number of samples from $\mathcal{N}(0,1)$ before seeing even one instance of the threshold being exceeded. On the other hand, for $X \sim \mathcal{N}(20,1)$, $\mathbf{P}[X \geq 20] = \frac{1}{2}$, so with high probability we will see the threshold exceeded roughly half the time. In that case, the approximation

$$\mathbf{E}[f(W)] \approx \frac{1}{n} \sum_{i=1}^{n} g(X_i), \qquad g(x) := f(x) e^{-\mu x + \frac{1}{2}\mu^2} \tag{4.1.8}$$

will already be good for a "reasonable" value of n, but with $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1)$ for the values of μ close to 20.

4.1.2 Removing a constant drift

We begin by considering the simplest case of Brownian motion with a constant drift, i.e., $X_t = W_t + \mu t$, $0 \leq t \leq T$, for some $\mu \in \mathbb{R}$. Suppose that we want to compute an expectation of the form $\mathbf{E}[f(X_{t_1}, X_{t_2}, \ldots, X_{t_n})]$, where the times $0 < t_1 < t_2 < \cdots < t_n \leq T$ are given. Let $q(x_1, \ldots, x_n)$ denote the probability density of the random vector $(X_{t_1}, \ldots, X_{t_n})$, so that

$$\mathbf{E}[f(X_{t_1}, X_{t_2}, \ldots, X_{t_n})] = \int_{\mathbb{R}^n} f(x_1, \ldots, x_n) q(x_1, \ldots, x_n) \, \mathrm{d}x_1 \ldots \mathrm{d}x_n. \tag{4.1.9}$$

Now, from the form of X_t we can write q as follows:

$$q(x_1, \ldots, x_n) = C \prod_{i=1}^{n} \exp\left(-\frac{(x_i - x_{i-1} - \mu(t_i - t_{i-t})^2}{2(t_i - t_{i-1})} \right), \tag{4.1.10}$$

where we have defined, as before, $(x_0, t_0) = (0, 0)$ and where

$$C = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi(t_i - t_{i-1})}}$$

is a normalization constant. Since

$$\frac{(x_i - x_{i-1} - \mu(t_i - t_{i-t})^2}{2(t_i - t_{i-1})} = \frac{(x_i - x_{i-1})^2}{2(t_i - t_{i-1})} - \mu(x_i - x_{i-1}) + \frac{\mu^2(t_i - t_{i-1})}{2}, \tag{4.1.11}$$

we can express q differently as

$$q(x_1, \ldots, x_n) = p(x_1, \ldots, x_n) \cdot \exp\left(\sum_{i=1}^{n} \left[\mu(x_i - x_{i-1}) - \frac{\mu^2(t_i - t_{i-1})}{2}\right]\right) \tag{4.1.12}$$

$$= p(x_1, \ldots, x_n) \exp\left(\mu x_n - \frac{\mu^2}{2} t_n\right), \tag{4.1.13}$$

where $p(x_1, \ldots, x_n) = C \prod_{i=1}^{n} \exp\left(-\frac{(x_i - x_{i-1})^2}{2(t_i - t_{i-1})}\right)$ is the probability density of $(W_{t_1}, \ldots, W_{t_n})$. Hence, we can write

$$\mathbf{E}[\varphi(X_{t_1}, \ldots, X_{t_n})] = \mathbf{E}\left[\varphi(W_{t_1}, \ldots, W_{t_n}) M_{t_n}\right], \tag{4.1.14}$$

where on the right-hand side the expectation is w.r.t. the probability law of the Brownian motion $(W_t)_{t\in[0,T]}$ and where the process $M_t := \exp(\mu W_t - \frac{\mu^2}{2} t)$ is a *martingale* w.r.t. the filtration $(\mathcal{F}_t)$ generated by the Brownian motion. The martingale property of (M_t) turns out to be rather important: Since $t_n \leq T$, we have $\mathbf{E}[M_T | \mathcal{F}_{t_n}] = M_{t_n}$, and therefore, using the law of iterated expectations twice, we get

$$\mathbf{E}\left[\varphi(W_{t_1}, \ldots, W_{t_n}) M_{t_n}\right] = \mathbf{E}[\varphi(W_{t_1}, \ldots, W_{t_n}) \mathbf{E}[M_T | \mathcal{F}_{t_n}]] \tag{4.1.15}$$

$$= \mathbf{E}[\mathbf{E}[\varphi(W_{t_1}, \ldots, W_{t_n}) M_T | \mathcal{F}_{t_n}]] \tag{4.1.16}$$

$$= \mathbf{E}\left[\varphi(W_{t_1}, \ldots, W_{t_n}) M_T\right]. \tag{4.1.17}$$

Thus, we have obtained a remarkable *change-of-measure formula*:

$$\mathbf{E}[\varphi(X_{t_1}, \ldots, X_{t_n})] = \mathbf{E}[\varphi(W_{t_1}, \ldots, W_{t_n}) M_T], \tag{4.1.18}$$

where on the left-hand side the expectation is w.r.t. $\boldsymbol{Q}$, the probability law of $(X_t)_{t\in[0,T]}$, while on the right-hand side the expectation is w.r.t. $\boldsymbol{P}$, the probability law of the Brownian motion $(W_t)_{t\in[0,T]}$. We can think of this as removing the drift from X_t; another difference is that, on the right-hand side, the "observable" $\varphi(\cdot)$ is multiplied by the exponential martingale M_T that depends neither on the choice of φ nor on the choice of the times $t_1, \ldots, t_n$. Hence, if we consider both processes as defined on the path space $(\Omega, \mathcal{B})$, then for any measurable function $f : \Omega \to \mathbb{R}$ of the entire continuous path $\omega = (\omega)_{t\in[0,T]}$, we can write

$$\mathbf{E}_{\boldsymbol{Q}}[f(\boldsymbol{X})] = \mathbf{E}_{\boldsymbol{P}}[f(\boldsymbol{W}) M_T]. \tag{4.1.19}$$

Since f is arbitrary, this gives us a relation between the two path-space measures $\boldsymbol{Q}$ and $\boldsymbol{P}$, one corresponding to Brownian motion with linear drift and one without: for any $A \in \mathcal{B}$,

$$\boldsymbol{Q}(A) = \mathbf{E}_{\boldsymbol{Q}}[\mathbf{1}_A] = \mathbf{E}_{\boldsymbol{P}}[\mathbf{1}_A M_T]. \tag{4.1.20}$$

4.1.3 A general Girsanov theorem

We start by introducing $(W_t)_{t\in[0,T]}$, a d-dimensional $\boldsymbol{P}$-Brownian motion, and take $(\mathcal{F}_t)_{t\in[0,T]}$ be the filtration generated by it.

Theorem 11. *Consider an Itô process*

$$X_t = \int_0^t F_s \,\mathrm{d}s + W_t, \qquad 0 \le t \le T, \tag{4.1.21}$$

where the dritf F_t, adapted to $\mathcal{F}_t$, is such that the process

$$M_t := \exp\left(-\int_0^t F_s^T \,\mathrm{d}W_s - \frac{1}{2}\int_0^t |F_s|^2 \,\mathrm{d}s \right) \tag{4.1.22}$$

satisfies $\mathbf{E}_{\boldsymbol{P}}[M_T] = 1$. If we let $\boldsymbol{Q}$ denote the path-space measure defined by $\mathbf{E}_{\boldsymbol{Q}}[\cdot] = \mathbf{E}_{\boldsymbol{P}}[\cdot M_T]$, then $(X_t)_{t\in[0,T]}$ is a $\boldsymbol{Q}$-Brownian motion.

Remark 6. *The condition $\mathbf{E}_{\boldsymbol{P}}[M_T] = 1$ ensures, among other things, that $\boldsymbol{Q}$ is indeed a probability measure on the path space $C([0,T];\mathbb{R}^d)$.*

Proof. We will assume $d = 1$ to keep things simple; the proof for $d > 1$ follows along the same lines. Using Itô's formula, we can write

$$\mathrm{d}X_t = F_t \,\mathrm{d}t + \mathrm{d}W_t \tag{4.1.23}$$

and

$$\mathrm{d}M_t = -M_t F_t \,\mathrm{d}W_t. \tag{4.1.24}$$

The first of these follows from the structure of X_t as an Itô process, while the second is readily verified by computing $\mathrm{d}\log M_t$:

$$\begin{aligned}
\mathrm{d}\log M_t &= \frac{\mathrm{d}M_t}{M_t} - \frac{1}{2M_t^2}\,\mathrm{d}[M]_t && (4.1.25)\\
&= -\frac{1}{M_t}\cdot M_t F_t \,\mathrm{d}W_t - \frac{1}{2M_t^2}\cdot |F_t|^2 M_t^2 \,\mathrm{d}t && (4.1.26)\\
&= -F_t \,\mathrm{d}W_t - \frac{1}{2}|F_t|^2 \,\mathrm{d}t && (4.1.27)
\end{aligned}$$

Then Itô's product rule gives

$$\begin{aligned}
\mathrm{d}(X_t M_t) &= X_t \,\mathrm{d}M_t + M_t \,\mathrm{d}X_t + \mathrm{d}X_t \,\mathrm{d}M_t && (4.1.28)\\
&= -X_t M_t F_t \,\mathrm{d}W_t + M_t(F_t \,\mathrm{d}t + \mathrm{d}W_t) - M_t F_t \,\mathrm{d}t && (4.1.29)\\
&= (1 - X_t F_t) M_t \,\mathrm{d}W_t. && (4.1.30)
\end{aligned}$$

Hence, both M_t and $X_t M_t$ are both $\mathcal{F}_t$-adapted Itô processes with zero drift, and are therefore both $\boldsymbol{P}$-martingales.[1] By contrast, since X_t has a nonzero drift, it is not a $\boldsymbol{P}$-martingale. We will now

[1] Strictly speaking, we have only shown that they are *local* martingales, but it can be proved that, under our assumptions, they are actually martingales.

show that this implies that X_t is a $\boldsymbol{Q}$-martingale. To that end, it suffices to show that, for any $T \geq t \geq s \geq 0$ and any event $A \in \mathcal{F}_s$,

$$\mathbf{E}_{\boldsymbol{Q}}[X_t \mathbf{1}_A] = \mathbf{E}_{\boldsymbol{Q}}[X_s \mathbf{1}_A]. \tag{4.1.31}$$

which is equivalent to $\mathbf{E}[X_t|\mathcal{F}_s] = X_s$. Using the definition of $\boldsymbol{Q}$, the law of iterated expectations, and the fact that both M_t and $X_t M_t$ are $\boldsymbol{P}$-martinagles, we can write

$$\begin{aligned}
\mathbf{E}_{\boldsymbol{Q}}[X_t \mathbf{1}_A] &= \mathbf{E}_{\boldsymbol{P}}[X_t M_T \mathbf{1}_A] && (4.1.32)\\
&= \mathbf{E}_{\boldsymbol{P}}[\mathbf{E}_{\boldsymbol{P}}[X_t M_T \mathbf{1}_A|\mathcal{F}_t]] && (4.1.33)\\
&= \mathbf{E}_{\boldsymbol{P}}[X_t \mathbf{E}_{\boldsymbol{P}}[M_T|\mathcal{F}_t]\mathbf{1}_A] && (4.1.34)\\
&= \mathbf{E}_{\boldsymbol{P}}[X_t M_t \mathbf{1}_A] && (4.1.35)\\
&= \mathbf{E}_{\boldsymbol{P}}[X_s M_s \mathbf{1}_A] && (4.1.36)\\
&= \mathbf{E}_{\boldsymbol{P}}[X_s \mathbf{E}_{\boldsymbol{P}}[M_T|\mathcal{F}_s]\mathbf{1}_A] && (4.1.37)\\
&= \mathbf{E}_{\boldsymbol{P}}[\mathbf{E}_{\boldsymbol{P}}[X_s M_T \mathbf{1}_A|\mathcal{F}_s]] && (4.1.38)\\
&= \mathbf{E}_{\boldsymbol{P}}[X_s M_T \mathbf{1}_A] && (4.1.39)
\end{aligned}$$

(a good exercise is to go through this chain of equalities and to identify which property is used where). Now, the last expression is equal to $\mathbf{E}_{\boldsymbol{Q}}[X_s \mathbf{1}_A]$ by the definition of $\boldsymbol{Q}$, so we are done.

Finally, observe that the process X_t has quadratic variation $[X]_t = t$, since the drift contributes nothing to the quadratic variation. Hence, under $\boldsymbol{Q}$, $(X_t)_{t\in[0,T]}$ is a martingale with continuous sample paths and $[X]_t = t$, and therefore it is a $\boldsymbol{Q}$-Brownian motion by Lévy's theorem. □

We can also give this result in a more symmetric form:

Theorem 12. *Let $\boldsymbol{P}$ denote the path-space probability law of the d-dimensional Itô process*

$$X_t = x + \int_0^t F_s \,\mathrm{d}s + \int_0^t G_s \,\mathrm{d}W_s, \qquad 0 \leq t \leq T, \tag{4.1.40}$$

where $x \in \mathbb{R}^d$ is a deterministic initial condition, (W_t) is a d-dimensional Brownian motion, and the random $d \times d$ matrices G_t are invertible. Furthermore, let $\tilde{F}_t$ be any adapted drift, such that the process $H_t := G_t^{-1}(F_t - \tilde{F}_t)$ is well-defined, and the process

$$M_t := \exp\left(-\int_0^t H_s^T \,\mathrm{d}W_s - \frac{1}{2}\int_0^t |H_s|^2 \,\mathrm{d}s\right) \tag{4.1.41}$$

satisfies $\mathbf{E}_{\boldsymbol{P}}[M_T] = 1$. Then for the path-space measure $\boldsymbol{Q}$ given by $\mathbf{E}_{\boldsymbol{Q}}[\cdot] = \mathbf{E}_{\boldsymbol{P}}[\cdot M_T]$, the process

$$\tilde{W}_t = W_t + \int_0^t H_s \,\mathrm{d}s \tag{4.1.42}$$

is a $\boldsymbol{Q}$-Brownian motion, and X_t can also be represented as

$$X_t = x + \int_0^t \tilde{F}_s \,\mathrm{d}s + \int_0^t G_s \,\mathrm{d}\tilde{W}_s, \qquad 0 \leq s \leq T. \tag{4.1.43}$$

Remark 7. *It is easy to see by considering $\tilde{F}_t = 0$ and $G_t = I$ that Theorem 11 is a special case of this.*

Proof. We already done most of the work in proving Theorem 11. In particular, it is easy to show, using essentially the same arguments, that M_t and $X_t M_t$ are both $\boldsymbol{P}$-martingales; the fact that $\tilde{W}_t$ is a $\boldsymbol{Q}$-Brownian motion is a consequence (in fact, the content) of Theorem 11. It only remains to verify that X_t can be represented as an Itô process using $\tilde{F}_t$, G_t, and the $\boldsymbol{Q}$-Brownian motion $\tilde{W}_t$. This is a routine computation:

$$\begin{aligned} \mathrm{d}X_t &= F_t\,\mathrm{d}t + G_t\,\mathrm{d}W_t && (4.1.44)\\ &= \tilde{F}_t\,\mathrm{d}t + G_t H_t\,\mathrm{d}t + \mathrm{d}W_t && (4.1.45)\\ &= \tilde{F}_t\,\mathrm{d}t + G_t(\mathrm{d}W_t + H_t\,\mathrm{d}t) && (4.1.46)\\ &= \tilde{F}_t\,\mathrm{d}t + G_t\,\mathrm{d}\tilde{W}_t, && (4.1.47) \end{aligned}$$

where the last line follows from the form of $\tilde{W}_t$. □

4.1.4 Absolute continuity

Hidden in the statement of Girsanov's theorem is a very important fact about probability laws of Itô processes: The probability laws of any two Itô processes that differ only in their drift terms are absolutely continuous w.r.t. each other. This fact is often used in computation of expectations. For example, consider a d-dimensional diffusion process $\boldsymbol{X} = (X_t)_{t\in[0,T]}$ given by

$$X_t = x + \int_0^t f(X_s)\,\mathrm{d}s + W_t \tag{4.1.48}$$

and let $\boldsymbol{Q}$ denote its probability law on the path space $\Omega = C([0,T];\mathbb{R}^d)$. Suppose we wanted to compute the expectation of some functional $h : \Omega \to \mathbb{R}$:

$$\mathbf{E}_{\boldsymbol{Q}}[h(\boldsymbol{X})] = \int_\Omega h(\boldsymbol{x})\boldsymbol{Q}(\mathrm{d}\boldsymbol{x}). \tag{4.1.49}$$

This type of "functional" integration arises in a variety of applications of stochastic calculus. The difficulty here, apart from the infinite-dimensional nature of the $\boldsymbol{x}$, is that we do not even have an explicit expression for $\boldsymbol{Q}$. However, with the help of Girsanov's theorem we can write this expectation as

$$\begin{aligned} \mathbf{E}_{\boldsymbol{Q}}[h(\boldsymbol{X})] &= \mathbf{E}_{\boldsymbol{P}}\left[h(x+\boldsymbol{W})\frac{\mathrm{d}\boldsymbol{Q}}{\mathrm{d}\boldsymbol{P}}(x+\boldsymbol{W})\right] && (4.1.50)\\ &= \mathbf{E}_{\boldsymbol{P}}\left[h(x+\boldsymbol{W})\exp\left(\int_0^T f(x+W_t)\,\mathrm{d}W_t - \frac{1}{2}\int_0^T |f(x+W_t)|^2\,\mathrm{d}t\right)\right] && (4.1.51) \end{aligned}$$

—this is still a functional integral, but now it involves "only" the probability law of the Brownian motion, and there are computational methods for approximating such functional integrals. To see how the above representation comes about, we simply apply Theorem 12 with $F_t = 0$, $\tilde{F}_t = f(X_t)$, $G_t = I$.

4.1.5 Weak solutions of SDEs

Consider an Itô SDE

$$\mathrm{d}X_t = f(X_t)\,\mathrm{d}t + \mathrm{d}W_t, \qquad 0 \le t \le T \tag{4.1.52}$$

Theorem 9 on the existence and uniqueness of solutions to SDEs showed that, under appropriate regularity conditions on the drift f, the above equation has a unique solution which can be expressed as a function of the initial condition X_0 and the *given* Brownian motion $(W_t)_{0\le t\le T}$ on a probability space $(\Omega, \mathcal{F}, \mathbf{P})$. This is known as a *strong solution*, and it has a clear system-theoretic interpretation: We can realize the process $(X_t)_{0\le t\le T}$ as the output of a system that takes the initial condition X_0 and the Brownian motion $(W_t)_{0\le t\le T}$ and produces the path $(X_t)_{0\le t\le T}$. Moreover, the system is causal in the sense that X_t depends only on X_0 and on $(W_s)_{0\le s\le t}$.

However, this is not the only sense in which an SDE like (4.1.52) can have a solution. For example, we can write it down and then ask whether there exists *some* probability space $(\Omega, \mathcal{F}, \mathbf{P})$, such that on that space we can construct both a Brownian motion $(W_t)_{0\le t\le T}$ and a process $(X_t)_{0\le t\le T}$, such that

$$X_t = \int_0^t f(X_s)\,\mathrm{d}s + W_t, \qquad 0 \le t \le T. \tag{4.1.53}$$

If such a construction is possible, then we say that (4.1.52) has a *weak solution*. It turns out that weak solutions exist whenever Girsanov theory applies. Let us take any probability space $(\Omega, \mathcal{F}, \mathbf{P})$, such that $(X_t)_{0\le t\le T}$ is a $\mathbf{P}$-*Brownian motion*. Now suppose that f is such that

$$\mathbf{E}_{\mathbf{P}}\left[\exp\left(\int_0^T f(X_t)^{\mathsf{T}}\,\mathrm{d}X_t - \frac{1}{2}\int_0^T |f(X_t)|^2\,\mathrm{d}t\right)\right] = 1, \tag{4.1.54}$$

and consider a new probability measure $\boldsymbol{Q}$ on $(\Omega, \mathcal{F})$, such that

$$\frac{\mathrm{d}\boldsymbol{Q}}{\mathrm{d}\boldsymbol{P}} = \exp\left(\int_0^T f(X_t)^{\mathsf{T}}\,\mathrm{d}X_t - \frac{1}{2}\int_0^T |f(X_t)|^2\,\mathrm{d}t\right). \tag{4.1.55}$$

Then, by Theorem 12, the process

$$W_t = X_t - \int_0^t f(X_s)\,\mathrm{d}s \tag{4.1.56}$$

is a $\boldsymbol{Q}$–Brownian motion, and thus we have constructed the pair (X_t, W_t), where $(W_t)_{0\le t\le T}$ is a Brownian motion and (4.1.53) holds. The catch with weak solutions is that, unlike the strong case where $(X_t)_{0\le t\le T}$ is generated causally from the given driving Brownian motion $(W_t)_{0\le t\le T}$, from (4.1.56) we see that the reverse is true: The Brownian motion $(W_t)_{0\le t\le T}$ is generated causally from the process $(X_t)_{0\le t\le T}$! Weak solutions are probabilistic in nature, unlike strong solutions that are closer in spirit to solutions of differential equations.

4.2 The Feynman–Kac formula

As we already saw in Chapter 2, there is a fundamental connection between diffusion processes and a certain type of second-order PDEs. For example, consider a function $u(x,t)$ defined by

$$u(x,t) := \frac{1}{(2\pi t)^{d/2}} \int_{\mathbb{R}^d} \varphi(x+\xi) e^{-|\xi|^2/2t} \, \mathrm{d}\xi, \qquad (x,t) \in \mathbb{R}^d \times [0,\infty) \tag{4.2.1}$$

where $\varphi : \mathbb{R}^d \to \mathbb{R}$ is a given C^2 function. On the one hand, it is a matter of straightforward calculus to show that it solves the PDE

$$\frac{\partial}{\partial t} u(x,t) = \frac{1}{2} \Delta u(x,t), \qquad u(x,0) = \varphi(x). \tag{4.2.2}$$

On the other, it also has a probabilistic representation

$$u(x,t) = \mathbf{E}[\varphi(x + W_t)], \tag{4.2.3}$$

where $(W_t)_{t \geq 0}$ is a standard d-dimensional Brownian motion. This, of course, is not a coincidence since $\mathcal{A} = \frac{1}{2}\Delta$ is the infinitesimal generator of the Brownian motion. More generally, consider a d-dimensional diffusion process $(X_t)_{t \geq 0}$ with drift $f(x)$ and diffusion matrix $a(x)$. Suppose that a function $u(x,t)$, which is C^2 in x and C^1 in t, solves the PDE

$$\frac{\partial}{\partial t} u(x,t) = \mathcal{A} u(x,t), \qquad u(x,0) = \varphi(x) \tag{4.2.4}$$

for a given C^2 function $\varphi : \mathbb{R}^d \to \mathbb{R}$, where

$$\mathcal{A} u(x,t) = f(x)^T \nabla u(x,t) + \frac{1}{2} \mathrm{tr}\{a(x) \nabla^2 u(x,t) \tag{4.2.5}$$

is the action of the infinitesimal generator $\mathcal{A}$ of $(X_t)_{t \geq 0}$ on $u(x,t)$ (as usual, ∇ and ∇^2 act on the space variable x, while keeping t fixed). Then it can be shown that it also has a probabilistic representation $u(x,t) = \mathbf{E}[\varphi(X_t) | X_0 = x]$. (In fact, the converse can also be shown to hold under additional regularity conditions.)

The relation between PDEs and diffusion processes can be pushed further. For example, let $u(x,t)$ be a solution of

$$\frac{\partial}{\partial t} u(x,t) = \mathcal{A} u(x,t) - \lambda u(x,t), \qquad u(x,0) = \varphi(x) \tag{4.2.6}$$

where $\lambda > 0$ is a fixed constant. Then $u(x,t) = e^{-\lambda t} \mathbf{E}[\varphi(X_t) | X_0 = x]$, where X_t, as before, is a diffusion process with generator $\mathcal{A}$. To show this, let $v(x,t) = e^{\lambda t}$. Then $v(x,0) = u(x,0) = \varphi(x)$, and

$$\frac{\partial}{\partial t} v(x,t) = \lambda v(x,t) + e^{\lambda t} \frac{\partial}{\partial t} u(x,t) \tag{4.2.7}$$

$$= \lambda v(x,t) + e^{\lambda t} \big(\mathcal{A} u(x,t) - \lambda u(x,t) \big) \tag{4.2.8}$$

$$= e^{\lambda t} \mathcal{A} u(x,t) \tag{4.2.9}$$

$$= \mathcal{A} v(x,t). \tag{4.2.10}$$

Thus, $u(x,t) = e^{-\lambda t} v(x,t) = e^{-\lambda t} \mathbf{E}[\varphi(X_t) | X_0 = x]$, as claimed. The following fundamental result, however, shows that we can obtain probabilistic representations of solutions of a much broader class of PDEs:

Theorem 13 (Feynman–Kac)**.** *Let $u(x,t)$ be a $C^{2,1}$ solution of the PDE*

$$\frac{\partial}{\partial t}u(x,t) = \mathcal{A}u(x,t) + V(x)u(x,t), \qquad u(x,0) = \varphi(x). \tag{4.2.11}$$

Then it has the following probabilistic representation:

$$u(x,t) = \mathbf{E}\left[\varphi(X_t)\exp\left(\int_0^t V(X_s)\,\mathrm{d}s\right)\middle| X_0 = x\right], \tag{4.2.12}$$

where $(X_t)_{t\geq 0}$ is a diffusion proess with infinitesimal generator $\mathcal{A}$.

Remark 8. *As will evident from the proof, the result also holds, with obvious modifications, when V depends on both x and t.*

Proof. Fix x and $t > 0$ and consider the following process $(M_s)_{0\leq s\leq t}$:

$$M_s := u(X_s, t-s)\exp\left(\int_0^s V(X_r)\,\mathrm{d}r\right). \tag{4.2.13}$$

Then $M_0 = u(X_0, t)$ and

$$M_t = u(X_t, 0)\exp\left(\int_0^t V(X_r)\,\mathrm{d}r\right) = \varphi(X_t)\exp\left(\int_0^t V(X_r)\,\mathrm{d}r\right). \tag{4.2.14}$$

We want to show that $(M_s)_{0\leq s\leq t}$ is a martingale, which would immediately imply that

$$u(x,t) = \mathbf{E}[M_0|X_0 = x] = \mathbf{E}[M_t|X_0 = x] = \mathbf{E}\left[\varphi(X_t)\exp\left(\int_0^t V(X_s)\,\mathrm{d}s\right)\middle| X_0 = x\right]. \tag{4.2.15}$$

Let $U_s := u(X_s, t-s)$ and $I_s := \exp\left(\int_0^s V(X_r)\,\mathrm{d}r\right)$. For the latter, we have $\mathrm{d}I_s = V(X_s)I_s\,\mathrm{d}s$. Furthermore, applying Itô's rule to $\varphi(x,s) = u(x, t-s)$ and using the fact that $u(x,t)$ solves (4.2.11) we obtain

$$\begin{aligned}
\mathrm{d}U_s &= -\dot{u}(X_s, t-s)\,\mathrm{d}s + \mathcal{A}u(X_s, t-s)\,\mathrm{d}s + \frac{\partial}{\partial x}u(X_s, t-s)g(X_s)\,\mathrm{d}W_s && (4.2.16)\\
&= -V(X_s)u(X_s, t-s)\,\mathrm{d}s + \frac{\partial}{\partial x}u(X_s, t-s)g(X_s)\,\mathrm{d}W_s && (4.2.17)\\
&= -V(X_s)U_s\,\mathrm{d}s + \frac{\partial}{\partial x}u(X_s, t-s)g(X_s)\,\mathrm{d}W_s. && (4.2.18)
\end{aligned}$$

Now, $M_s = U_sI_s$, so Itô's product rule gives

$$\begin{aligned}
\mathrm{d}M_s &= U_s\,\mathrm{d}I_s + I_s\,\mathrm{d}U_s && (4.2.19)\\
&= U_sV(X_s)I_s\,\mathrm{d}s + I_s\left(-V(X_s)U_s\,\mathrm{d}s + \frac{\partial}{\partial x}u(X_s, t-s)g(X_s)\,\mathrm{d}W_s\right) && (4.2.20)\\
&= I_s\frac{\partial}{\partial x}u(X_s, t-s)g(X_s)\,\mathrm{d}W_s, && (4.2.21)
\end{aligned}$$

which shows that M_s is indeed a martingale. □

4.2.1 A killing interpretation

Let $(X_t)_{t\geq 0}$ be a continuous-time d-dimensional Markov process, and let τ be a nonnegative-valued random variable. We then define a Markov process $(\bar{X}_t)_{t\geq 0}$ *killed at time* τ by adding a special "death" state $\mathfrak{d}$ and taking

$$\bar{X}_t := \begin{cases} X_t, & t < \tau \\ \mathfrak{d}, & t \geq \tau \end{cases} \tag{4.2.22}$$

In other words, the path of $\bar{X}_t$ is obtained by following the path of X_t until the "killing time" τ, at which point the process enters the death state and remains there. Given a function $\varphi : \mathbb{R}^d \to \mathbb{R}$, we extend it by taking $\varphi(\mathfrak{d}) = 0$. Then

$$\mathbf{E}[\varphi(\bar{X}_t)] = \mathbf{E}[\varphi(X_t)\mathbf{1}_{\{\tau > t\}}]. \tag{4.2.23}$$

For example, suppose $\tau \sim \mathrm{Exp}(\lambda)$, i.e., τ has exponential distribution with parameter λ and is independent of (X_t). Then

$$\mathbf{E}[\varphi(\bar{X}_t)] = \mathbf{E}[\varphi(X_t)]\mathbf{P}[\tau > t] \tag{4.2.24}$$

$$= e^{-\lambda t}\mathbf{E}[\varphi(X_t)]. \tag{4.2.25}$$

We can consider other choices of τ that depend on the path of X_t. For instance, let a nonnegative function $V : \mathbb{R}^d \to \mathbb{R}$ be given and take

$$\tau := \inf\left\{ t \geq 0 : \int_0^t V(X_s)\,\mathrm{d}s \geq \gamma \right\}, \tag{4.2.26}$$

where $\gamma \sim \mathrm{Exp}(1)$ is independent of (X_t). Then, letting $(\mathcal{F}_t)$ be the filtration generated by (X_t), we have

$$\mathbf{E}[\varphi(\bar{X}_t)] = \mathbf{E}\left[\varphi(X_t)\mathbf{1}_{\{\tau > t\}}\right] \tag{4.2.27}$$

$$= \mathbf{E}\left[\varphi(X_t)\mathbf{1}_{\{\gamma > \int_0^t V(X_s)\,\mathrm{d}s\}}\right] \tag{4.2.28}$$

$$= \mathbf{E}\left[\mathbf{E}\left[\varphi(X_t)\mathbf{1}_{\{\gamma > \int_0^t V(X_s)\,\mathrm{d}s\}}\middle|\mathcal{F}_t\right]\right] \tag{4.2.29}$$

$$= \mathbf{E}\left[\varphi(X_t)\mathbf{E}\left[\mathbf{1}_{\{\gamma > \int_0^t V(X_s)\,\mathrm{d}s\}}\middle|\mathcal{F}_t\right]\right] \tag{4.2.30}$$

$$= \mathbf{E}\left[\varphi(X_t)\exp\left(-\int_0^t V(X_s)\,\mathrm{d}s\right)\right]. \tag{4.2.31}$$

Hence, if $(X_t)_{t\geq 0}$ is a diffusion process with deterministic initial condition $X_0 = x$, then the Feynman–Kac formula tells us that the solution of the PDE

$$\frac{\partial}{\partial t}u(x,t) = \mathcal{A}u(x,t) - V(x)u(x,t), \qquad u(x,0) = \varphi(x) \tag{4.2.32}$$

is equal to the expectation $\mathbf{E}[\varphi(\bar{X}_t)|X_0 = x]$, where $\bar{X}_t$ is the process X_t killed at time τ defined in (4.2.26). We can think of killing as a variant of importance sampling, where the exponential weight $\exp(-\int_0^t V(X_s)\,\mathrm{d}s)$ is used to reject (or "kill") those paths of (X_t) that incur a large total cost at time t, with the running cost measured by the potential function V.

4.3 Problems

1. Recall that Girsanov's theorem says the following: If $(X_t)_{0\le t\le 1}$ is a d-dimensional $\boldsymbol{P}$-Brownian motion, then, for any d-dimensional adapted process $(F_t)_{0\le t\le 1}$ such that

$$\mathbf{E}_{\boldsymbol{P}}\left[\exp\left(\int_0^1 F_t^T\,\mathrm{d}X_t - \frac{1}{2}\int_0^1 |F_t|^2\,\mathrm{d}t\right)\right] = 1 \tag{4.3.1}$$

the process

$$W_t := X_t - \int_0^t F_s\,\mathrm{d}s, \qquad 0 \le t \le 1 \tag{4.3.2}$$

is a $\boldsymbol{Q}$-Brownian motion, where $\boldsymbol{Q}$ is the probability measure on the path space $\Omega = C([0,1];\mathbb{R}^d)$ defined by $\mathbf{E}_{\boldsymbol{Q}}[\cdot] = \mathbf{E}_{\boldsymbol{P}}[\cdot M_T]$ with

$$M_T = \exp\left(\int_0^1 F_t^T\,\mathrm{d}X_t - \frac{1}{2}\int_0^1 |F_t|^2\,\mathrm{d}t\right). \tag{4.3.3}$$

For any deterministic vector $v \in \mathbb{R}^d$, let

$$M_T^v := \exp\left(\int_0^1 (F_t + v)^T\,\mathrm{d}X_t - \frac{1}{2}\int_0^1 |F_t + v|^2\,\mathrm{d}t\right), \tag{4.3.4}$$

so that $M_T = M_T^0$. Using martingale methods and Lévy's characterization of Brownian motion, it can be shown that

$$\mathbf{E}_{\mathbf{P}}[M_T^v] = 1,\ \forall v \in \mathbb{R}^d \quad\Longrightarrow\quad W_t = X_t - \int_0^t F_s\,\mathrm{d}s \text{ is a } \boldsymbol{Q}\text{-Brownian motion.} \tag{4.3.5}$$

In this problem, we will prove a converse result: If $(F_t)_{0\le t\le 1}$ is an adapted process such that the process W_t defined in (4.3.2) is a $\boldsymbol{Q}$-Brownian motion with $\boldsymbol{Q}$ defined in (4.3.3), then (4.3.7) holds.

(i) Let $v \in \mathbb{R}^d$ be a deterministic vector. Prove that

$$\log M_T^v = \log M_T + v^T W_1 - \frac{1}{2}|v|^2. \tag{4.3.6}$$

(ii) Use the result of (i) to prove that

$$\mathbf{E}_{\boldsymbol{P}}[M_T^v] = 1, \qquad \forall v \in \mathbb{R}^d. \tag{4.3.7}$$

Hint: Recall that, under $\boldsymbol{Q}$, $W_1 \sim \mathcal{N}(0, I_d)$.

2. Let $(X_t)_{0\le t\le 1}$, $X_0 = 0$, be a one-dimensional diffusion process governed by the SDE

$$\mathrm{d}X_t = f(X_t)\,\mathrm{d}t + \mathrm{d}W_t, \tag{4.3.8}$$

where the drift $f : \mathbb{R} \to \mathbb{R}$ satisfies the nonlinear differential equation

$$f'(x) + f^2(x) = ax^2 + bx + c \tag{4.3.9}$$

for $a \ge 0$ and arbitrary b, c. Prove that, for any bounded measurable function $\varphi : \mathbb{R} \to \mathbb{R}$,

$$\mathbf{E}[\varphi(X_1)] = e^{-c/2}\mathbf{E}\left[\varphi(W_1)\exp\left(F(W_1) - \frac{1}{2}\int_0^1 (aW_t^2 + bW_t)\,\mathrm{d}t\right)\right], \tag{4.3.10}$$

where $F(x) := \int_0^x f(y)\,\mathrm{d}y$, and where on the right-hand side $(W_t)_{0\le t\le 1}$ is a standard one-dimensional Brownian motion.

Hint: Use Girsanov's theorem to remove the drift from X_t, then apply Itô's rule to $F(W_t)$.

3. Let f, F, a, b, c be as in Problem 2. Let $u(x,t)$ be a $C^{2,1}$ solution of the PDE

$$\frac{\partial}{\partial t}u(x,t) = \frac{1}{2}\frac{\partial^2}{\partial x^2}u(x,t) - \frac{1}{2}(ax^2 + bx)u(x,t), \qquad u(x,0) = e^{F(x)}, \tag{4.3.11}$$

Show that $u(0,1) = e^{c/2}$.

4.4 Notes and further reading

My presentation of Girsanov's theory, including the motivation through importance sampling, largely follows that of [Ste01]. Importance sampling is a rich and interesting subject in its own right; see [Buc04] for a clean exposition geared toward engineering system simulation. The first result on shifting the drift from a Brownian motion via a change of measure was obtained by Cameron and Martin [CM44] for deterministic drifts; the paper of Girsanov [Gir60] generalized it to arbitrary Itô processes. The martingale proof of Girsanov's theorem, which I got from Steele's book, seems to originate in the work of Beneš [Ben71], although this paper is not cited by Steele. The paper of Van Schuppen and Wong [SW74] gives what is probably the most general Girsanov-type result. A brief discussion of the system-theoretic interpretation of strong solutions of SDEs can be found in Section 5.2 of [KS98].

The origins of the Feynman–Kac formula go back to [Kac49]. As Kac recounts in his delightful autobiography *Enigmas of Chance* [Kac85], he attented a physics colloquium talk by Richard Feynman in 1947, when both of them were at Cornell University. Feynman's talk was about the new approach to quantum mechanics he had developed, based on "sums over histories" or what is now referred to as "Feynman path integrals" [FH65]. Feynman's use of his approach to derive a fundamental quantum-mechanical quantity known as the *propagator*, which is related to solutions of Schrödinger's equation. As Kac remembers it [Kac85, p. 116], the steps in Feynman's derivation were reminiscent of some arguments pertaining to the connection between Brownian motion and the heat equation. See [Roe94, Sim05] for detailed expositions of the path (or functional) integral approach.

Part II

Applications

Chapter 5

Stochastic control and estimation

Control, in very broad strokes, is about shaping the behavior of some system in order to attain a given goal. For example, if we have a continuous-time deterministic system with input $u(t) \in \mathbb{R}^m$, state $x(t) \in \mathbb{R}^n$, and output $y(t) \in \mathbb{R}^p$ described by the state-space model

$$\dot{x}(t) = f(x(t), u(t)), \tag{5.0.1a}$$
$$y(t) = h(x(t)) \tag{5.0.1b}$$

then the following are all examples of control:

1. If we have perfect observation of the state, i.e., $y(t) = h(x(t))$, then we may seek a control law $u : [0, 1] \to \mathbb{R}^m$ to transfer the system's state from a given initial point $x(0) = x_0$ to a given final point $x(1) = x_1$. (We say that we want to *control* the system from x_0 to x_1) in unit time.) Whether this is at all possible depends on the dynamical law of the system, i.e., on $f(\cdot, \cdot)$; this is a question of *controllability*.

2. If at least one such control law exists, we then may want to ask whether the transfer can be accomplished in the best possible way—thus, we may want to minimize the *control effort*

$$\frac{1}{2} \int_0^1 |u(t)|^2 \, \mathrm{d}t \tag{5.0.2}$$

   over all controls $u : [0, 1] \to \mathbb{R}^m$ that accomplish the transfer from x_0 to x_1 in unit time. This is a instance of *optimal control*.

3. More generally, we may want to choose a control law u to minimize a total cost of the form

$$\int_0^1 q(x(t), u(t)) \, \mathrm{d}t + r(x(1)) \tag{5.0.3}$$

   over all "well-behaved" control laws $u : [0, 1] \to \mathbb{R}^m$, where q is the *instantaneous cost* and r is the *terminal cost*. The interpretation of this is that, for a given initial condition x_0, we choose $u(\cdot)$ to trade off the terminal cost $r(x(1))$ against the integrated cost of "getting there" from $x(0) = x_0$. This can be done using state feedback control, i.e., there exists a function $k : \mathbb{R}^n \times [0, 1] \to \mathbb{R}^m$ such that the optimal control law $u(t)$ is given by $k(x(t), t)$, where $x(t)$ evolves according to the closed-loop dynamics $\dot{x}(t) = f(x(t), k(x(t)), t)$. However, due

to the nature of solutions of deterministic ODEs, we can also look for optimal control laws in open-loop form, i.e., just as functions of time. These are two complementary viewpoints, formalized by dynamic programming and the maximum principle, respectively.

4. Even more generally, we may want to accomplish various goals while only having access to a processed version of the state, i.e., the sensor output $y(t) = h(x(t))$. In this case, we may still want to choose $u : [0, T] \to \mathbb{R}^m$ to minimize the total cost as in the above example, but with the additional restriction that the control $u(t)$ can only be chosen on the basis of the output trajectory $(y(s) : 0 \le s < t)$. This is the problem optimal control with *partial observations*, and it is intimately connected with such concepts as *observability*, i.e., when the initial state $x(0)$ can be recovered from an observation trajectory up to time t.

The solutions to these problems are all known if the system in question is *linear*, i.e., if $f(x, u) = Ax + Bu$ and $g(x) = Cx$ for some matrices A, B, C of appropriate shapes. (The systems don't have to be time-invariant, i.e., the system matrices may all depend on t.) For nonlinear systems, there is no general theory, and in fact some questions can be answered in general only locally, e.g., we may want to ask whether a given point x_0 has an open neighborhood K, such that it is possible to transfer the system from x_0 to any given point $x_1 \in K$ in finite time. Nevertheless, we see that there are a number of basic questions regarding control and estimation that one may want to pose about dynamical systems.

Our interest here lies with stochastic systems modeled by SDEs, so some of the above questions (such as controllability or observability) do not have obvious stochastic counterparts. Thus, we will focus for the most part on the optimal control problem. In that context, we can also consider the cases of full vs. partial observations, and this will force us to consider some estimation problems as well. We will start with the latter.

5.1 Linear estimation: the Kalman–Bucy filter

We consider the following system of linear SDEs with time-varying coefficients:

$$\mathrm{d}X_t = A(t)X_t\,\mathrm{d}t + B(t)\,\mathrm{d}W_t \tag{5.1.1}$$

$$\mathrm{d}Y_t = C(t)X_t\,\mathrm{d}t + D(t)\,\mathrm{d}V_t \tag{5.1.2}$$

where $(W_t)_{t\ge0}$ and $(V_t)_{t\ge0}$ are two independent Brownian motions that are also independent of the initial condition X_0. We assume that X_0 is Gaussian with mean m_0 and covariance matrix K_0 and that $Y_0 = 0$. Thus, both X_t and Y_t are Gaussian processes. We also assume that, for each t, the matrix $D(t)D(t)^T$ is invertible. For each t, let $\mathcal{F}_t$ be the σ-algebra generated by $(X_s, Y_s)_{0\le s\le t}$ and let $\mathcal{F}_t^Y$ be the observation σ-algebra generated by the observations $(Y_s)_{0\le s\le t}$ only. Our goal is to show that the conditional mean $\hat{X}_t := \mathbf{E}[X_t|\mathcal{F}_t^Y]$ is determined by a linear SDE with initial condition $\hat{X}_0 = m_0$.

We start by recalling a fundamental result on minimum mean-square error estimation in the jointly Gaussian setting:

Lemma 1. *Let I be a parameter set, let $(V_\alpha)_{\alpha\in I}$ be a collection of random variables indexed by the elements of I, and let $\mathcal{V}$ be the σ-algebra generated by the V_α's. Let U be a random variable, such that U and $(V_\alpha)_{\alpha\in A}$ are jointly Gaussian. Then the conditional mean $\hat{U} = \mathbf{E}[U|\mathcal{V}]$ is uniquely characterized by the following properties:*

- *measurability*—$\hat{U}$ *is a measurable function of* $(V_\alpha)_{\alpha\in I}$*;*
- *unbiasedness*—$\mathbf{E}[U-\hat{U}]=0$*;*
- *orthogonality*—$\mathbf{E}[(U-\hat{U})V_\alpha]=0$ *for all* $\alpha\in I$*.*

We will look for $\hat{X}_t$ defined through a linear SDE of the form

$$\mathrm{d}\hat{X}_t = A(t)\hat{X}_t\,\mathrm{d}t + L(t)(\mathrm{d}Y_t - C(t)\hat{X}_t\,\mathrm{d}t) \tag{5.1.3}$$

with initial condition $\hat{X}_0 = m_0$, where $L(t)$ is a matrix that we need to determine. Since the solution of (5.1.3) has the form

$$\hat{X}_t = m_0 + \int_0^t \big(A(s)-L(s)C(s)\big)\hat{X}_s\,\mathrm{d}s + \int_0^t L(s)\,\mathrm{d}Y_s, \tag{5.1.4}$$

we see that $\hat{X}_t$ is a measurable function of $(Y_s)_{0\le s\le t}$. Moreover, if we define the error process $\tilde{X}_t := X_t - \hat{X}_t$, then $\mathbf{E}[\tilde{X}_0]=0$ and

$$\mathrm{d}\tilde{X}_t = \big(A(t)-L(t)C(t)\big)\tilde{X}_t\,\mathrm{d}t + B(t)\,\mathrm{d}W_t - L(t)D(t)\,\mathrm{d}V_t, \tag{5.1.5}$$

which gives $\mathbf{E}[X_t-\hat{X}_t]=\mathbf{E}[\tilde{X}_t]=0$ for all t. Thus, $\hat{X}_t$ satisfies the measurability and the unbiasedness requirements of the lemma with $U=X_t$, $I=[0,t]$, and $\mathcal{V}=\mathcal{F}_t^Y$. Finally, we would like to show orthogonality. To that end, let us also define the process $\nu_t := Y_t - \int_0^t C(s)\hat{X}_s\,\mathrm{d}s$ with the initial condition $\nu_0=0$. Then the joint process $(\tilde{X}_t,\nu_t)$ is Gaussian since it has the initial condition $\tilde{X}_0\sim\mathcal{N}(0,K_0)$, $\nu_0=0$, and is governed by the system of linear SDEs

$$\mathrm{d}\tilde{X}_t = \big(A(t)-L(t)C(t)\big)\tilde{X}_t\,\mathrm{d}t + B(t)\,\mathrm{d}W_t - L(t)D(t)\,\mathrm{d}V_t \tag{5.1.6}$$

$$\mathrm{d}\nu_t = C(t)\tilde{X}_t\,\mathrm{d}t + D(t)\,\mathrm{d}V_t \tag{5.1.7}$$

or, in more suggestive matrix notation,

$$\begin{pmatrix}\mathrm{d}\tilde{X}_t\\ \mathrm{d}\nu_t\end{pmatrix} = \begin{pmatrix}A(t)-L(t)C(t) & 0\\ C(t) & 0\end{pmatrix}\begin{pmatrix}\tilde{X}_t\\ \nu_t\end{pmatrix}\mathrm{d}t + \begin{pmatrix}B(t) & -L(t)D(t)\\ 0 & D(t)\end{pmatrix}\begin{pmatrix}\mathrm{d}W_t\\ \mathrm{d}V_t\end{pmatrix}. \tag{5.1.8}$$

From this, it is not hard to write down the differential equations for the covariance and the cross-covariance matrices $K_{\tilde{X}}(t) := \mathbf{E}[\tilde{X}_t\tilde{X}_t^T]$, $K_{\tilde{X}\nu}(t) := \mathbf{E}[\tilde{X}_t\nu_t^T]$, and $K_\nu(t) := \mathbf{E}[\nu_t\nu_t^T]$:

$$\begin{aligned}\dot{K}_{\tilde{X}}(t) &= K_{\tilde{X}}(t)\big(A(t)-L(t)C(t)\big)^T + \big(A(t)-L(t)C(t)\big)K_{\tilde{X}}(t)\\ &\qquad + B(t)B(t)^T + L(t)D(t)D(t)^TL(t)^T \qquad (5.1.9)\\ \dot{K}_{\tilde{X}\nu}(t) &= \big(A(t)-L(t)C(t)\big)K_{\tilde{X}\nu}(t)^T + K_{\tilde{X}}(t)C(t)^T - L(t)D(t)D(t)^T \qquad (5.1.10)\\ \dot{K}_\nu(t) &= C(t)K_{\tilde{X}\nu}(t)^T + K_{\tilde{X}\nu}(t)^TC(t)^T + D(t)D(t)^T \qquad (5.1.11)\end{aligned}$$

with the initial conditions $K_{\tilde{X}}(0)=K_0$, $K_{\tilde{X}\nu}(0)=0$, $K_\nu(0)=0$. The choice of

$$L(t) = K_{\tilde{X}}(t)C(t)^T\big(D(t)D(t)^T\big)^{-1} \tag{5.1.12}$$

is rather fortuitous, since it simultaneously results in $K_{\tilde{X}\nu}(t) = 0$ for all t,

$$K_\nu(t) = \int_0^t D(s)D(s)^T \,\mathrm{d}s, \tag{5.1.13}$$

and

$$\dot{K}_{\tilde{X}}(t) = K_{\tilde{X}}(t)A(t)^T + A(t)K_{\tilde{X}}(t) - K_{\tilde{X}}(t)C(t)^T\big(D(t)D(t)^T\big)^{-1}C(t)K_{\tilde{X}}(t) + B(t)B(t)^T, \tag{5.1.14}$$

where the latter is the *Riccati differential equation* with the initial condition $K_{\tilde{X}}(0) = K_0$. Thus, if in (5.1.3) we take $L(t)$ according to (5.1.12), where the error covariance matrix $K_{\tilde{X}}(t)$ is determined from (5.1.14), then we see that, for each t, the error $\tilde{X}_t = X_t - \hat{X}_t$ has zero mean and is orthogonal to ν_t:

$$\mathbf{E}[\tilde{X}_t\nu_t^T] = K_{\tilde{X}\nu}(t) = 0. \tag{5.1.15}$$

From here, it is not hard to show that $\tilde{X}_t$ is orthogonal to all ν_s, $0 \le s \le t$. Thus, by Lemma 1, $\hat{X}_t = \mathbf{E}[X_t|\mathcal{F}_t^\nu]$, where

$$\nu_t = Y_t - \int_0^t C(s)\hat{X}_s \,\mathrm{d}s \tag{5.1.16}$$

is the *innovations process* with covariance matrix given by (5.1.13). Recall, however, that we were after the conditional mean of X_t given $\mathcal{F}_t^Y$, not $\mathcal{F}_t^\nu$, but it turns out that these two conditional means are equal since the observations $(Y_s)_{0\le s\le t}$ and the innovations $(\nu_s)_{0\le s\le t}$ generate *the same σ-algebra:*

$$\mathcal{F}_t^Y = \mathcal{F}_t^\nu. \tag{5.1.17}$$

This result, known as *causal equivalence*, follows from the fact that $\hat{X}_t$ can be expressed in two ways as

$$\hat{X}_t = \hat{X}_0 + \int_0^t A(s)\hat{X}_s \,\mathrm{d}s + \int_0^t L(s)\,\mathrm{d}\nu_s \tag{5.1.18}$$

$$= \hat{X}_0 + \int_0^t \big(A(s) - L(s)C(s)\big)\hat{X}_s \,\mathrm{d}s + \int_0^t L(s)\,\mathrm{d}Y_s, \tag{5.1.19}$$

which, in combination with (5.1.16), shows that we can generate the innovations trajectory $(\nu_s)_{0\le s\le t}$ from $(Y_s)_{0\le s\le t}$ and vice versa. Thus, $\hat{X}_t$ is the conditional mean $\mathbf{E}[X_t|\mathcal{F}_t^Y]$.

The innovations process ν_t has a number of useful properties. For instance, it has independent increments, i.e., for all $0 \le t_1 \le t_2 \le t_3 \le t_4$, the random vectors $\nu_{t_4} - \nu_{t_3}$ and $\nu_{t_2} - \nu_{t_1}$ are independent. Moreover, for $t \ge s \ge 0$, the increments $\nu_t - \nu_s$ are independent of $\mathcal{F}_s^Y$. To see this, observe that, for $t \ge s$,

$$\nu_t - \nu_s = Y_t - Y_s - \int_s^t C(r)\hat{X}_r \,\mathrm{d}r \tag{5.1.20}$$

$$= \int_s^t C(r)\tilde{X}_r \,\mathrm{d}r + \int_s^t D(r)\,\mathrm{d}V_r, \tag{5.1.21}$$

and both $(\tilde{X}_r)_{r\geq s}$ and $(\tilde{V}_r)_{r\geq s}$ are independent of $\mathcal{F}_s^Y$. The independent increments property of ν_t is proved similarly. Moreover, we can compute the covariance matrix of $\nu_t - \nu_s$:

$$\mathbf{E}[(\nu_t - \nu_s)(\nu_t - \nu_s)^T] = \int_s^t D(r)D(r)^T \,\mathrm{d}r. \tag{5.1.22}$$

This can be seen as follows: For any deterministic vector $v \in \mathbb{R}^p$ (where p is the dimension of Y_t), define the process $Z_t^v := v^T(\nu_t - \nu_s)$. Then

$$\mathrm{d}Z_t^v = v^T C(t)\tilde{X}_t \,\mathrm{d}t + v^T D(t)\,\mathrm{d}V_t \tag{5.1.23}$$

so we can easily compute the joint quadratic variation $\mathrm{d}[Z^v, Z^w]_t = v^T D(t)D(t)^T w \,\mathrm{d}t$. Itô's product rule then gives

$$\begin{aligned}\mathrm{d}(Z_t^v Z_t^w) &= Z_t^v \,\mathrm{d}Z_t^w + Z^w \,\mathrm{d}Z_t^v + \mathrm{d}[Z^v, Z^w]_t &(5.1.24)\\ &= Z_t^v\big(w^T C(t)\tilde{X}_t \,\mathrm{d}t + w^T D(t)\,\mathrm{d}V_t\big) + Z_t^w\big(v^T C(t)\tilde{X}_t \,\mathrm{d}t + v^T D(t)\,\mathrm{d}V_t\big) + v^T D(t)D(t)^T w \,\mathrm{d}t. &(5.1.25)\end{aligned}$$

Since Z_t^v and $\tilde{X}_t$ are independent, taking expectations gives

$$\mathbf{E}[v^T(\nu_t - \nu_s)(\nu_t - \nu_s)^T w] = \int_s^t v^T D(r)D(r)^T w \,\mathrm{d}r. \tag{5.1.26}$$

Since v and w are arbitrary, we get (5.1.13). Now, since $D(t)D(t)^T$ is invertible, the process $\hat{W}_t := \sqrt{(D(t)D(t)^T)^{-1}}\nu_t$ has independent increments and $\mathrm{d}[\hat{W}^i, \hat{W}^j]_t = \delta_{ij}\,\mathrm{d}t$ for all $1 \leq i, j \leq p$, so it is a standard p-dimensional Brownian motion adapted to $\mathcal{F}_t^\nu$ (and hence to $\mathcal{F}_t^Y$). We can therefore express the filter process $\hat{X}_t$ as

$$\hat{X}_t = m_0 + \int_0^t A(s)\hat{X}_s \,\mathrm{d}s + \int_0^t L(s)\sqrt{D(s)D(s)^\top}\,\mathrm{d}\hat{W}_s. \tag{5.1.27}$$

In particular, if $D(t)$ is symmetric and positive definite (and thus the observation noise V_t is a standard p-dimensional Brownian motion), then the observation process Y_t obeys the SDE

$$\mathrm{d}Y_t = C(t)\hat{X}_t \,\mathrm{d}t + D(t)\,\mathrm{d}\hat{W}_t, \tag{5.1.28}$$

where $D(t)\,\mathrm{d}\hat{W}_t = \mathrm{d}\nu_t$ provides the new information, on top of $C(t)\hat{X}_t \,\mathrm{d}t$.

5.2 Nonlinear filtering

A more general problem involves observations of the form

$$Y_t = \int_0^t h(X_s)\,\mathrm{d}s + V_t, \qquad 0 \leq t \leq T, \tag{5.2.1}$$

where (V_t) is a standard Brownian motion and where (X_t) is the n-dimensional signal (or state) process governed by the SDE

$$\mathrm{d}X_t = f(X_t)\,\mathrm{d}t + g(X_t)\,\mathrm{d}W_t. \tag{5.2.2}$$

Here, (W_t) is a standard n-dimensional Brownian motion independent of the initial condition X_0. In what follows, we will impose the following assumptions:

1. The observation process (Y_t) is scalar.

2. The signal process (X_t) and the measurement function h are such that

$$\mathbf{E}\left[\int_0^T h^2(X_t)\,\mathrm{d}t\right] < \infty. \tag{5.2.3}$$

3. The Brownian motion (V_t) is independent of the signal process (X_t).

The first assumption is introduced mainly for convenience; in the multidimensional case, we can do all the calculations on a per-coordinate basis. The third assumption can also be relaxed, but at the expense of considerably more technical arguments. It is also possible to consider observation noise of the form $\int_0^t g(s)\,\mathrm{d}V_s$ with $g > 0$ everywhere.

Let $\mathcal{F}_t^Y := \sigma(Y_s : 0 \le s \le t)$ denote the σ-algebra generated by the observations up to time t. The problem of *nonlinear filtering* is to characterize, for each t, the conditional probability law of X_t given $\mathcal{F}_t^Y$. To be more precise, let a function $\varphi : \mathbb{R}^n \to \mathbb{R}$ be given, and let $\pi_t(\varphi)$ denote the conditional expectation

$$\pi_t(\varphi) := \mathbf{E}[\varphi(X_t)|\mathcal{F}_t^Y]. \tag{5.2.4}$$

The goal is to recursively compute $\pi_t(\varphi)$ for any sufficiently regular (say, bounded and continuous) function φ. There are two approaches to this problem, the innovations approach and the reference measure approach.

5.2.1 The innovations approach

The innovations approach was first proposed by Kailath in the linear case and then extended by Frost and Kailath to the nonlinear setting. Given any process (ξ_t), let $\hat{\xi}_t := \mathbf{E}[\xi_t|\mathcal{F}_t^Y]$. In particular, let $Z_t := h(X_t)$ and define the innovations process

$$\nu_t := Y_t - \int_0^t \hat{Z}_s\,\mathrm{d}s. \tag{5.2.5}$$

Just as in the Kalman–Bucy setting, the increments $\nu_{t+h} - \nu_t$ represent the "new" information about the process (Z_t) contained in the observations Y_s for times s between t and $t+h$. In fact, just like in the linear case, we have the following:

Lemma 2. *The innovations process (ν_t) is an $\mathcal{F}_t^Y$-Brownian motion.*

Proof. The innovations process (ν_t) is an $\mathcal{F}_t^Y$-martingale:

$$\begin{aligned}
\mathbf{E}[\mathrm{d}\nu_t|\mathcal{F}_t^Y] &= \mathbf{E}[\mathrm{d}Y_t - \hat{Z}_t\,\mathrm{d}t|\mathcal{F}_t^Y] && (5.2.6)\\
&= \mathbf{E}[(Z_t - \hat{Z}_t)\,\mathrm{d}t + \mathrm{d}V_t|\mathcal{F}_t^Y] && (5.2.7)\\
&= \mathbf{E}[Z_t - \hat{Z}_t|\mathcal{F}_t^Y]\,\mathrm{d}t + \mathbf{E}[\mathrm{d}V_t|\mathcal{F}_t^Y], && (5.2.8)
\end{aligned}$$

where the first conditional expectation vanishes by the definition of $\hat{Z}_t$, while the second one vanishes since $\mathrm{d}V_t$ is zero-mean and independent of $\mathcal{F}_t^Y$. Moreover, using (5.2.1) and the definition of ν_t in (5.2.5) can express (Y_t) in two ways as follows:

$$Y_t = \int_0^t Z_s\,\mathrm{d}s + V_t = \int_0^t \hat{Z}_s\,\mathrm{d}s + \nu_t. \tag{5.2.9}$$

Hence, the quadratic variation $[\nu]_t$ is equal to the quadratic variation $[Y]_t$, which is in turn equal to t. Thus, since (ν_t) is an $\mathcal{F}_t^Y$-martingale with continuous sample paths and $[\nu]_t = t$, it is an $\mathcal{F}_t^Y$-Brownian motion by Lévy's theorem. □

We will also make use of the following martingale representation result:

Lemma 3 ([FKK72]). *Every square-integrable $\mathcal{F}_t^Y$-martingale M_t can be represented as*

$$M_t = \mathbf{E}[M_0] + \int_0^t \eta_s \, \mathrm{d}\nu_s. \tag{5.2.10}$$

Next, we give a representation for the conditional mean $\hat{\xi}_t = \mathbf{E}[\xi_t | \mathcal{F}_t^Y]$, where ξ_t is a process of the form

$$\xi_t = \xi_0 + \int_0^t F_s \, \mathrm{d}s + M_t, \tag{5.2.11}$$

where M_t is an $\mathcal{F}_t$-martingale, where $\mathcal{F}_t$ is the σ-algebra generated by $(X_s, Y_s)_{0 \le s \le t}$. Then we will specialize it to the case $\xi_t = \varphi(X_t)$.

Theorem 14. *Suppose that the joint quadratic variation $[M, V]_t$ between the martingale M_t in (5.2.11) and the Brownian motion V_t in (5.2.1) vanishes. Then $\hat{\xi}_t$ admits the following representation:*

$$\hat{\xi}_t = \hat{\xi}_0 + \int_0^t \hat{F}_s \, \mathrm{d}s + \int_0^t \left(\widehat{\xi_s Z_s} - \hat{\xi}_s \hat{Z}_s \right) \mathrm{d}\nu_s. \tag{5.2.12}$$

Proof. First, we show that the process

$$\bar{M}_t := \hat{\xi}_t - \hat{\xi}_0 - \int_0^t \hat{F}_s \, \mathrm{d}s \tag{5.2.13}$$

is a zero-mean $\mathcal{F}_t^Y$-martingale. Indeed, $\bar{M}_0 = 0$, and

$$\begin{aligned} \mathbf{E}[\mathrm{d}\hat{\xi}_t | \mathcal{F}_t^Y] &= \mathbf{E}[\mathrm{d}\xi_t | \mathcal{F}_t^Y] && (5.2.14) \\ &= \mathbf{E}[F_t | \mathcal{F}_t^Y] \, \mathrm{d}t + \mathbf{E}[\mathrm{d}M_t | \mathcal{F}_t^Y] && (5.2.15) \\ &= \mathbf{E}[\hat{F}_t | \mathcal{F}_t^Y] \, \mathrm{d}t, && (5.2.16) \end{aligned}$$

where in the last step we have used the fact that M_t is a $\mathcal{F}_t^Y$-martingale. Rearranging shows that $\mathbf{E}[\mathrm{d}\bar{M}_t | \mathcal{F}_t^Y] = 0$, as claimed. We can now invoke Lemma 3 to conclude that there exists a $\mathcal{F}_t^Y$-adapted process (η_t), such that

$$\hat{\xi}_t = \hat{\xi}_0 + \int_0^t \hat{F}_s \, \mathrm{d}s + \int_0^t \eta_s \, \mathrm{d}\nu_s. \tag{5.2.17}$$

It remains to identify the process η_t. We start with the observation that

$$\mathbf{E}[(\xi_t - \hat{\xi}_t) Y_t | \mathcal{F}_t^Y] = 0, \tag{5.2.18}$$

which follows from the properties of conditional expectation. For $\xi_t Y_t$, Itô's product rule gives

$$\xi_t Y_t = \xi_0 Y_0 + \int_0^t \xi_s \, \mathrm{d}Y_s + \int_0^t Y_s \, \mathrm{d}\xi_s + \int_0^t \mathrm{d}[\xi, Y]_s \tag{5.2.19}$$

$$= \xi_0 Y_0 + \int_0^t \xi_s (Z_s \, \mathrm{d}s + \mathrm{d}V_s) + \int_0^t Y_s (F_s \, \mathrm{d}s + \mathrm{d}M_s), \tag{5.2.20}$$

where we have also used the fact that the quadratic variation term vanishes since $[V, M]_t = 0$ by hypothesis. Analogously, for $\hat{\xi}_t Y_t$,

$$\hat{\xi}_t Y_t = \hat{\xi}_0 Y_0 + \int_0^t \hat{\xi}_s \, \mathrm{d}Y_s + \int_0^t Y_s \, \mathrm{d}\hat{\xi}_s + \int_0^t \mathrm{d}[\hat{\xi}, Y]_s \tag{5.2.21}$$

$$= \xi_0 Y_0 + \int_0^t \hat{\xi}_s (Z_s \, \mathrm{d}s + \mathrm{d}V_s) + \int_0^t Y_s (\hat{F}_s \, \mathrm{d}s + \eta_s \, \mathrm{d}\nu_s) + \int_0^t \eta_s \, \mathrm{d}s, \tag{5.2.22}$$

where we have used the (easily established) fact that $\mathrm{d}[\hat{\xi}, Y]_t = \eta_t \, \mathrm{d}t$. Taking conditional expectations given $\mathcal{F}_t^Y$, we obtain

$$\mathbf{E}[\xi_t Y_t | \mathcal{F}_t^Y] = \int_0^t \mathbf{E}[\xi_s Z_s | \mathcal{F}_s^Y] \, \mathrm{d}s + \int_0^t \mathbf{E}[Y_s F_s | \mathcal{F}_s^Y] \, \mathrm{d}s \tag{5.2.23}$$

$$= \int_0^t \widehat{\xi_s Z_s} \, \mathrm{d}s + \int_0^t Y_s \hat{F}_s \, \mathrm{d}s \tag{5.2.24}$$

and

$$\mathbf{E}[\hat{\xi}_t Y_t | \mathcal{F}_t^Y] = \int_0^t \mathbf{E}[\hat{\xi}_s Z_s] \, \mathrm{d}s + \int_0^t \mathbf{E}[Y_s \hat{F}_s | \mathcal{F}_s^Y] \, \mathrm{d}s + \int_0^t \eta_s \, \mathrm{d}s \tag{5.2.25}$$

$$= \int_0^t \hat{\xi}_s \hat{Z}_s \, \mathrm{d}s + \int_0^t Y_s \hat{F}_s \, \mathrm{d}s + \int_0^t \eta_s \, \mathrm{d}s. \tag{5.2.26}$$

Plugging these expressions into (5.2.18) shows that $\eta_t = \widehat{\xi_t Z_t} - \hat{\xi}_t \hat{Z}_t$. □

Let us now use Theorem 14 to derive the nonlinear filter. Thus, let $\xi_t = \varphi(X_t)$, for a C^2 function φ. Then

$$\hat{\xi}_t = \mathbf{E}[\varphi(X_t) | \mathcal{F}_t^Y] = \pi_t(\varphi). \tag{5.2.27}$$

With this in hand, we have the following:

Theorem 15. *The nonlinear filter $\pi_t(\varphi)$ admits the following representation:*

$$\pi_t(\varphi) = \pi_0(\varphi) + \int_0^t \pi_s(\mathcal{A}\varphi) \, \mathrm{d}s + \int_0^t \big(\pi_s(h\varphi) - \pi_s(h)\pi_s(\varphi)\big) \, \mathrm{d}\nu_s, \tag{5.2.28}$$

where $\pi_0(\varphi) = \mathbf{E}[\varphi(X_0)]$, $\mathcal{A}$ is the infinitesimal generator of the signal diffusion process, and $h\varphi$ denotes the function $x \mapsto h(x)\varphi(x)$.

Proof. By Itô's rule, the process $\xi_t = \varphi(X_t)$ can be expressed as

$$\xi_t = \xi_0 + \int_0^t \mathcal{A}\varphi(X_s)\,\mathrm{d}s + \int_0^t \nabla\varphi(X_s)^T g(X_s)\,\mathrm{d}W_s, \tag{5.2.29}$$

This has the form (5.2.11) with $F_s = \mathcal{A}\varphi(X_s)$ and $\mathrm{d}M_s = \nabla\varphi(X_s)^T g(X_s)\,\mathrm{d}W_s$. Moreover, Lemma 4.2 in [FKK72] shows that $[M, V]_t = 0$ follows from the independence of (X_t) and (V_t). Hence, we can apply Theorem 14, and (5.2.28) follows since, e.g., $\hat{Z}_t = \mathbf{E}[h(X_t)|\mathcal{F}_t^Y] = \pi_t(h)$, etc. □

The expression (5.2.28) for the conditinoal expectation $\pi_t(\varphi)$ is known as the *Kushner–Stratonovich equation*. It is a stochastic *partial* differential equation due to the presence of the generator $\mathcal{A}$ in the right-hand side. Even though it gives the solution of the nonlinear filtering problem, it does not lead to a finite-dimensional recursive representation apart from some special cases. To see why this is so, let us consider the case when the signal process (X_t) is scalar and we wish to compute the conditional mean of X_t given $\mathcal{F}_t^Y$, i.e., $\hat{X}_t = \mathbf{E}[X_t|\mathcal{F}_t^Y]$, which is equal to $\pi_t(\varphi)$ with $\varphi(x) = x$. Then, as can be seen fron (5.2.28), we need to have expressions for $\pi_t(\mathcal{A}\varphi) = \pi_t(f) = \widehat{f(X_t)}$, for $\pi_t(h) = \hat{Z}_t = \widehat{h(X_t)}$, and for $\pi_t(h\varphi) = \widehat{Z_t X_t} = \widehat{h(X_t)X_t}$, and each of these, in turn, comes with its own Kushner–Stratonovich equation. This is known as the *closure problem*.

Example 8 (The Kalman–Bucy filter)**.** *Suppose that the signal process (X_t) and the observation process (Y_t) are both scalar and given by*

$$X_t = X_0 + \int_0^t aX_s\,\mathrm{d}s + bW_t \tag{5.2.30}$$

$$Y_t = \int_0^t cX_s\,\mathrm{d}s + V_t \tag{5.2.31}$$

where $X_0 \sim \mathcal{N}(m_0, \sigma^2)$, and we are interested in computing $\hat{X}_t = \mathbf{E}[X_t|\mathcal{F}_t^Y]$. Of course, we have already worked out the more general, multidimensional and time-varying, version of this problem; this example is meant to illustrate that the Kalman–Bucy filter is a special instance of the nonlinear filter that does not suffer from the closure problem.

The Kushner–Stratonovich equation for $\hat{X}_t$ takes the form

$$\hat{X}_t = \mathbf{E}[X_0] + \int_0^t a\hat{X}_s\,\mathrm{d}s + \int_0^t c(\widehat{X_s^2} - (\hat{X}_s)^2)\,\mathrm{d}\nu_s \tag{5.2.32}$$

$$= \mathbf{E}[X_0] + \int_0^t a\hat{X}_s\,\mathrm{d}s + \int_0^t c(\widehat{X_s^2} - (\hat{X}_s)^2)(\mathrm{d}Y_s - c\hat{X}_s\,\mathrm{d}s) \tag{5.2.33}$$

$$= \mathbf{E}[X_0] + \int_0^t a\hat{X}_s\,\mathrm{d}s + \int_0^t cK_t(\mathrm{d}Y_s - c\hat{X}_s\,\mathrm{d}s), \tag{5.2.34}$$

where $K_t := \mathbf{E}[(X_t - \hat{X}_t)^2|\mathcal{F}_t^Y]$ is the conditional error variance. Since (X_t, Y_t) is a Gaussian process, the error covariance K_t is nonrandom, and can be precomputed by solving the ODE

$$\frac{\mathrm{d}}{\mathrm{d}t}K_t = 2aK_t - c^2K_t^2 + b^2, \tag{5.2.35}$$

which is just the scalar Riccati equation.

5.2.2 The reference measure approach

The alternative approach to nonlinear filtering starts from the seemingly trivial observation that, when $h \equiv 0$ (i.e., when the path $(Y_s)_{0\le s\le t}$ is *completely uninformative* about the signal X_t), $\pi_t(\varphi) = \mathbf{E}[\varphi(X_t)]$ for all t. More generally, for any measurable function F of the signal path $\boldsymbol{X}_t := (X_s)_{0\le s\le t}$ and the observation path $\boldsymbol{Y}_t := (Y_s)_{0\le s\le t}$,

$$\mathbf{E}[F(\boldsymbol{X}_t, \boldsymbol{Y}_t)|\mathcal{F}_t^Y] = \int F(\boldsymbol{x}_t, \boldsymbol{Y}_t)\boldsymbol{P}_{\boldsymbol{X}}(\mathrm{d}\boldsymbol{x}_t), \tag{5.2.36}$$

that is, we simply marginalize out the signal process. But this is something we can arrange by making a Girsanov change of measure!

To that end, let $\boldsymbol{P}$ denote the probability law of $\boldsymbol{X}_T$ and $\boldsymbol{Y}_T$, and define a new probability law $\tilde{\boldsymbol{P}}$ by

$$\frac{\mathrm{d}\boldsymbol{P}}{\mathrm{d}\tilde{\boldsymbol{P}}} := \exp\left(\int_0^T Z_t\,\mathrm{d}Y_t - \frac{1}{2}\int_0^T |Z_t|^2\,\mathrm{d}t\right), \tag{5.2.37}$$

where, as we recall, $Z_t = h(X_t)$. Under $\tilde{\boldsymbol{P}}$, the signal process $\boldsymbol{X}_T$ has the same marginal probability law as under $\boldsymbol{P}$, but the observation process $\boldsymbol{Y}_T$ is now a Brownian motion *independent* of $\boldsymbol{X}_T$. Thus, we find ourselves in the situation described in the preceding paragraph, and we will now capitalize on this. First, let us define, for each $t \in [0,T]$, the following functional of the paths $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$:

$$\Lambda_t(\boldsymbol{X}_t, \boldsymbol{Y}_t) := \exp\left(\int_0^t h(X_s)\,\mathrm{d}Y_s - \frac{1}{2}\int_0^t |h(X_s)|^2\,\mathrm{d}s\right). \tag{5.2.38}$$

We will now prove the following result, known as the *Kallianpur–Striebel formula*:

$$\pi_t(\varphi) = \frac{\int \varphi(x_t)\Lambda_t(\boldsymbol{x}_t, \boldsymbol{Y}_t)\boldsymbol{P}_{\boldsymbol{X}_t}(\mathrm{d}\boldsymbol{x}_t)}{\int \Lambda_t(\boldsymbol{x}_t, \boldsymbol{Y}_t)\boldsymbol{P}_{\boldsymbol{X}_t}(\mathrm{d}\boldsymbol{x}_t)} \tag{5.2.39}$$

The starting point is the *conditional Bayes theorem*: Let μ and ν be two probablity measures on a common probability space $(\Omega, \mathcal{F})$ that are mutually absolutely continuous w.r.t. each other, and let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra. Then, for any bounded random variable X,

$$\mathbf{E}_\mu[X|\mathcal{G}] = \frac{\mathbf{E}_\nu[X\frac{\mathrm{d}\mu}{\mathrm{d}\nu}|\mathcal{G}]}{\mathbf{E}_\nu[\frac{\mathrm{d}\mu}{\mathrm{d}\nu}|\mathcal{G}]}, \tag{5.2.40}$$

where $\mathbf{E}[\cdot]$ (respectively, $\mathbf{E}_\nu[\cdot]$) denotes expectation w.r.t. μ (respectively, ν). To prove (5.2.40), let

A be an arbitrary event in $\mathcal{G}$. Then, using the properties of conditional expectations, we can write

$$\mathbf{E}_\nu\left[\mathbf{1}_A\mathbf{E}_\nu\left[X\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\middle|\mathcal{G}\right]\right] = \mathbf{E}_\nu\left[\mathbf{E}_\nu\left[\mathbf{1}_A X\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\middle|\mathcal{G}\right]\right] \tag{5.2.41}$$

$$= \mathbf{E}_\nu\left[\mathbf{1}_A X\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right] \tag{5.2.42}$$

$$= \mathbf{E}_\mu[\mathbf{1}_A X] \tag{5.2.43}$$

$$= \mathbf{E}_\mu[\mathbf{1}_A\mathbf{E}_\mu[X|\mathcal{G}]] \tag{5.2.44}$$

$$= \mathbf{E}_\nu\left[\mathbf{1}_A\mathbf{E}_\mu[X|\mathcal{G}]\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right] \tag{5.2.45}$$

$$= \mathbf{E}_\nu\left[\mathbf{1}_A\mathbf{E}_\mu[X|\mathcal{G}]\mathbf{E}_\nu\left[\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\middle|\mathcal{G}\right]\right]. \tag{5.2.46}$$

Since A was arbitrary, we conclude that

$$\mathbf{E}_\nu\left[X\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\middle|\mathcal{G}\right] = \mathbf{E}_\mu[X|\mathcal{G}]\mathbf{E}_\nu\left[\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\middle|\mathcal{G}\right], \tag{5.2.47}$$

and then (5.2.40) follows since $\mathbf{E}_\nu\left[\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\middle|\mathcal{G}\right]$ is almost surely positive by absolute continuity. To obtain (5.2.39), we apply (5.2.40) to $\mu = \boldsymbol{P}$, $\nu = \tilde{\boldsymbol{P}}$, $\mathcal{G} = \mathcal{F}_t^Y$, and $X = \varphi(X_t)$. Here, we also make use of the fact that Λ_t is a martingale under $\tilde{\boldsymbol{P}}$ (which is, again, a consequene of Girsanov's theorem), so that (denoting by $\tilde{\mathbf{E}}[\cdot]$ the expectation w.r.t. $\tilde{\boldsymbol{P}}$)

$$\tilde{\mathbf{E}}[\varphi(X_t)\Lambda_T(\boldsymbol{X}_T,\boldsymbol{Y}_T)|\mathcal{F}_t^Y] = \tilde{\mathbf{E}}[\varphi(X_t)\Lambda_t(\boldsymbol{X}_t,\boldsymbol{Y}_t)|\mathcal{F}_t^Y]. \tag{5.2.48}$$

The Kallianpur–Striebel formula is often expressed in terms of the *unnormalized filter* σ_t, which is defined as

$$\sigma_t(\varphi) := \tilde{\mathbf{E}}[\varphi(X_t)\Lambda_t|\mathcal{F}_t^Y] \tag{5.2.49}$$

$$= \int \varphi(x_t)\Lambda_t(\boldsymbol{x}_t,\boldsymbol{Y}_t)\boldsymbol{P}_{\boldsymbol{X}_t}(\mathrm{d}\boldsymbol{x}_t), \tag{5.2.50}$$

where φ is any bounded measurable function. This is simply the quantity in the numerator of (5.2.39), but the denominator is also of this form with the constant function $\varphi(x) \equiv 1$. That is,

$$\pi_t(\varphi) = \frac{\sigma_t(\varphi)}{\sigma_t(1)}. \tag{5.2.51}$$

We can also obtain a recursive representation for the unnormalized filter, known as the *Duncan–Mortensen–Zakai equation*. Assume φ is C^2. First, by Itô's rule

$$\mathrm{d}\varphi(X_t) = \mathcal{A}\varphi(X_t)\,\mathrm{d}t + \nabla\varphi(X_t)^T g(X_t)\,\mathrm{d}W_t; \tag{5.2.52}$$

moreover, $\mathrm{d}\Lambda_t = \Lambda_t Z_t\,\mathrm{d}Y_t$, and the joint quadratic variation of $\varphi(X_t)$ and Λ_t vanishes since the Brownian motion processes (W_t) and (V_t) are independent. Therefore, Itô's product rule gives

$$\mathrm{d}(\varphi(X_t)\Lambda_t) = \varphi(X_t)\,\mathrm{d}\Lambda_t + \Lambda_t\,\mathrm{d}\varphi(X_t) + \mathrm{d}\varphi(X_t)\,\mathrm{d}\Lambda_t \tag{5.2.53}$$

$$= \varphi(X_t)\Lambda_t Z_t\,\mathrm{d}Y_t + \Lambda_t\mathcal{A}\varphi(X_t)\,\mathrm{d}t + \Lambda_t\nabla\varphi(X_t)^T g(X_t)\,\mathrm{d}W_t. \tag{5.2.54}$$

Integrating, we get

$$\varphi(X_t)\Lambda_t = \varphi(X_0) + \int_0^t \mathcal{A}\varphi(X_s)\Lambda_s \,\mathrm{d}s + \int_0^t \varphi(X_s)Z_s\Lambda_s \,\mathrm{d}Y_s \tag{5.2.55}$$

$$= \varphi(X_0) + \int_0^t \mathcal{A}\varphi(X_s)\Lambda_s \,\mathrm{d}s + \int_0^t \varphi(X_s)h(X_s)\Lambda_s \,\mathrm{d}Y_s. \tag{5.2.56}$$

Taking expectations w.r.t. $\tilde{\mathbf{E}}$ conditioned on $\mathcal{F}_t^Y$ and using the definition of σ_t yields

$$\sigma_t(\varphi) = \sigma_0(\varphi) + \int_0^t \sigma_s(\mathcal{A}\varphi) \,\mathrm{d}s + \int_0^t \sigma_s(h\varphi) \,\mathrm{d}Y_s, \tag{5.2.57}$$

where, just as before, $h\varphi$ is shorthand for the function $x \mapsto h(x)\varphi(x)$. This is the Duncan–Mortensen–Zakai equation, and from there one can derive the Kushner–Stratonovich equation (5.2.28) by applying Itô's product rule to $\frac{\sigma_t(\varphi)}{\sigma_t(1)}$.

5.3 Optimal control of diffusion processes

Now that we have discussed the estimation and filtering problems, we move on to the problem of optimal control. First, we need to develop a suitable stochastic analogue of the controlled deterministic system (5.0.1). For the most part, we will consider the case of complete observations. This suggests replacing the deterministic state dynamics (5.0.1a) with an SDE of the form

$$\mathrm{d}X_t = f(X_t, u_t) \,\mathrm{d}t + g(X_t, u_t) \,\mathrm{d}W_t, \tag{5.3.1}$$

where we now have the n-dimensional state process $(X_t)_{0\le t\le 1}$, the m-dimensional input (or control) process $(u_t)_{0\le t\le 1}$, and a k-dimensional Brownian motion $(W_t)_{0\le t\le 1}$. Note that the control input may, in principle, affect both the drift and the diffusion matrix.

Now, we have to be careful about specifying in which sense, if any, the SDE (5.3.1) has a solution. This will, in turn, necessitate imposing some restrictions on the admissible control processes. The most obvious, direct approach is to require that the functions $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ and $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^{n\times k}$ satisfy

$$|f(x,u) - f(x',u')| + \|g(x,u) - g(x',u')\| \le K(|x - x'| + |u - u'|) \tag{5.3.2}$$

and

$$|f(x,u)|^2 + \|g(x,u)\|^2 \le K(1 + |x|^2 + |u|^2) \tag{5.3.3}$$

for some constant $0 \le K < \infty$, for all $x, x' \in \mathbb{R}^n$ and all $u, u' \in \mathbb{R}^m$. Then we can take a time-varying state feedback control law $u = k(x,t)$ for any function $k : \mathbb{R}^n \times [0,1] \to \mathbb{R}^m$ satisfying

$$|k(x,t) - k(x',t)| \le K'|x - x'| \tag{5.3.4}$$

and

$$|k(x,t)|^2 \le K'(1 + |x|^2) \tag{5.3.5}$$

for some constant $0 \le K' < \infty$, for all $x, x' \in \mathbb{R}^n$ and all $t \in [0,1]$. Under these assumptions, the functions $f^k(x,t) := f(x,k(x,t))$ and $g^k(x,t) := g(x,k(x,t))$ will satisfy

$$\begin{aligned} |f^k(x,t) - f^k(x',t)| &= |f(x,k(x,t)) - f(x',k(x',t))| && (5.3.6)\\ &\le K(|x-x'| + |k(x,t) - k(x',t)|) && (5.3.7)\\ &\le K''|x-x'| && (5.3.8) \end{aligned}$$

and

$$\begin{aligned} |f^k(x,t)|^2 + \|g^k(x,t)\|^2 &= |f(x,k(x,t))|^2 + \|g(x,k(x,t))\|^2 && (5.3.9)\\ &\le K(1 + |x|^2 + |k(x,t)|^2) && (5.3.10)\\ &\le \tilde{K}(1+|x|^2) && (5.3.11) \end{aligned}$$

for some other constant $\tilde{K}$, for all $x, x' \in \mathbb{R}^n$ and all $t \in [0,1]$. Under these assumptions, the SDE

$$\mathrm{d}X_t^k = f^k(X_t^k, t)\,\mathrm{d}t + g^k(X_t^k, t)\,\mathrm{d}W_t \tag{5.3.12}$$

will have a unique strong solution $(X_t^k)_{0\le t\le 1}$ for each initial condition $X_0 = x_0$, and we can therefore take $X_t = X_t^k$ and $u_t = k(X_t^k)$ in (5.3.1). Of course, these are fairly severe restrictions, and we may often settle for the existence of *weak* solutions (see Section 4.1.5) that are unique in probability law. That is, for each $x \in \mathbb{R}^n$, there exists some probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{0\le t\le 1}, \boldsymbol{P})$, a $\boldsymbol{P}$-Brownian motion $(W_t)_{0\le t\le 1}$ adapted to the filtration $(\mathcal{F}_t)_{0\le t\le 1}$, and two $\mathcal{F}_t$-adapted processes $(X_t)_{0\le t\le 1}$ and $(u_t)_{0\le t\le 1}$, such that (5.3.1) has a solution with $X_0 = x$, i.e.,

$$X_t = x + \int_0^t f(X_s, u_s)\,\mathrm{d}s + \int_0^t g(X_s, u_s)\,\mathrm{d}W_s, \qquad 0 \le t \le 1. \tag{5.3.13}$$

Moreover, the probability law of $(X_t, u_t)_{0\le t\le 1}$ on the path space $\Omega = C([0,1]; \mathbb{R}^n \times \mathbb{R}^m)$ is unique. The adaptedness requirement is entirely natural as it expresses that the controls are *causal*, i.e., do not make use of any information from the future. The uniqueness of the probability law of (X_t, u_t) is needed to ensure that the expected cost is well-defined. When the control is of state feedback form, i.e., $u = k(x,t)$ for some $k : \mathbb{R}^n \times [0,1] \to \mathbb{R}^m$, the above discussion simply means that the SDE (5.3.12) has a weak solution unique in probability law.

With all of this in mind, we will simply say that there is a class $\mathbb{U}$ of *admissible input* (or *control*) *processes* $\boldsymbol{u} = (u_t)_{0\le t\le 1}$, such that the controlled SDE (5.3.1) has a solution, in an appropriate sense, for each initial condiotion $X_0 = x$. We can then define, for each admissible control process $\boldsymbol{u}$, each $x \in \mathbb{R}^n$, and each $t \in [0,1]$, the *expected cost-to-go*

$$J(x,t;\boldsymbol{u}) := \mathbf{E}\left[\int_t^1 q(X_s, u_s)\,\mathrm{d}s + r(X_1)\middle| X_t = x\right] \tag{5.3.14}$$

and the *value function*

$$V(x,t) := \min_{\boldsymbol{u}\in\mathbb{U}} J(x,t;\boldsymbol{u}). \tag{5.3.15}$$

We say that $\bar{\boldsymbol{u}} \in \mathbb{U}$ is *optimal* if it achieves the value function, i.e., if $J(x,t;\bar{\boldsymbol{u}}) = V(x,t)$ for all x, t. Just like in the deterministic case, the computation of the value function relies on a *dynamic*

programming argument. The underlying idea, known as *Bellman's optimality principle*, is as follows: If $\boldsymbol{u} = (u_t)_{0\le t\le 1}$ is an optimal control process, then $(u_s)_{t\le s\le 1}$ is optimal on the time interval $[t,1]$, so replacing the trajectory $(u_s)_{t\le s\le 1}$ with some other admissible trajectory $(\tilde{u}_s)_{t\le s\le 1}$ will result in larger overall expected cost. This idea can be used as the basis for a rigorous derivation of a certain PDE that the value function V should satisfy, known as the *Hamilton–Jacobi–Bellman equation*, given in (5.3.17) below. In general, solving the HJB equation is a difficult task, although in some cases we can guess a solution based on some structural considerations (we will see examples of this later on). If our guess turns out to be correct, then we have obtained an optimal control. This is the content of the so-called *verification theorem*.

First, some notation. To each fixed $u \in \mathbb{R}^m$, we will associate a second-order linear differential operator $\mathcal{A}^u$ defined by

$$\mathcal{A}^u\varphi(x,t) := f(x,u)^T\nabla\varphi(x,t) + \frac{1}{2}\mathrm{tr}\{g(x,u)g(x,u)^T\nabla^2\varphi(x,t)\}. \tag{5.3.16}$$

We can think of $\mathcal{A}^u$ as the infinitesimal generator of a diffusion process with drift $f(\cdot,u)$ and diffusion matrix $g(\cdot,u)g(\cdot,u)^T$.

Theorem 16. *Suppose that $V : \mathbb{R}^n \times [0,1] \to \mathbb{R}$ is a $C^{2,1}$ solution of the Hamilton–Jacobi–Bellman equation*

$$\frac{\partial}{\partial t}V(x,t) + \min_{u\in\mathbb{R}^m}\left\{\mathcal{A}^uV(x,t) + q(x,u)\right\} = 0, \qquad (x,t)\in\mathbb{R}^n\times[0,1] \tag{5.3.17}$$

subject to the terminal condition $V(x,1) = r(x)$. Then we have the following:

1. *For any admissible control $\tilde{\boldsymbol{u}}$, any x, and any t,*

$$J(x,t;\tilde{\boldsymbol{u}}) \ge V(x,t). \tag{5.3.18}$$

2. *If there exists a function $k : \mathbb{R}^n \times [0,1] \to \mathbb{R}^m$, such that*

$$\mathcal{A}^{k(x,t)}V(x,t) + q(x,k(x,t)) = \min_{u\in\mathbb{R}^m}\left\{\mathcal{A}^uV(x,t) + q(x,u)\right\}, \qquad \forall (x,t)\in\mathbb{R}^n\times[0,1] \tag{5.3.19}$$

and the control $\bar{\boldsymbol{u}}$ corresponding to the state feedback law $\bar{u} = k(x,t)$ is admissible, then $J(x,t;\bar{\boldsymbol{u}}) = V(x,t)$ for all $(x,t)\in\mathbb{R}^n\times[0,1]$, and therefore $\bar{\boldsymbol{u}}$ is optimal.

Proof. For the first item, let $\tilde{\boldsymbol{u}}\in\mathbb{U}$, $x\in\mathbb{R}^n$, and $t\in[0,1]$ be given. Let $(\tilde{X}_s)_{t\le s\le 1}$ be a solution of

$$\mathrm{d}\tilde{X}_s = f(\tilde{X}_s,\tilde{u}_s)\,\mathrm{d}s + g(\tilde{X}_s,\tilde{u}_s)\,\mathrm{d}W_s, \qquad t\le s\le 1 \tag{5.3.20}$$

with $\tilde{X}_t = x$. By Itô's rule,

$$V(\tilde{X}_1,1) = V(\tilde{X}_t,t) + \int_0^t\left(\dot{V}(\tilde{X}_s,s) + \mathcal{A}^{\tilde{u}_s}V(\tilde{X}_s,s)\right)\mathrm{d}s + M_{t,1} \tag{5.3.21}$$

$$= V(x,t) + \int_0^t\left(\dot{V}(\tilde{X}_s,s) + \mathcal{A}^{\tilde{u}_s}V(\tilde{X}_s,s)\right)\mathrm{d}s + M_{t,1}, \tag{5.3.22}$$

where $M_{t,1} = \int_t^1 \nabla V(\tilde{X}_s,\tilde{u}_s)^T g(\tilde{X}_s,\tilde{u}_s)\,\mathrm{d}W_s$ has zero mean. Since V solves the HJB equation (5.3.17),

$$\dot{V}(x,s) + \mathcal{A}^uV(x,s) \ge -q(x,u), \qquad \forall x,u,s. \tag{5.3.23}$$

Therefore,

$$\int_0^t \left(\dot{V}(\tilde{X}_s, s) + \mathcal{A}^{\tilde{u}_s} V(\tilde{X}_s, s)\right) \mathrm{d}s \geq -\int_0^t q(\tilde{X}_s, \tilde{u}_s)\, \mathrm{d}s. \tag{5.3.24}$$

Since $V(\tilde{X}_1, 1) = r(\tilde{X}_1)$, we have

$$\mathbf{E}[r(\tilde{X}_1)|\tilde{X}_t = x] \geq V(x,t) - \mathbf{E}\left[\int_t^1 q(\tilde{X}_s, \tilde{u}_s)\, \mathrm{d}s \middle| \tilde{X}_t = x\right], \tag{5.3.25}$$

so we get (5.3.18) after rearranging terms. The same argument can be used to prove the second item since now

$$\dot{V}(x,s) + \mathcal{A}^{k(x,s)} V(x,s) = -q(x, k(x,s)), \qquad \forall x, s \tag{5.3.26}$$

by hypothesis on k. □

The verification theorem gives a sufficient condition for optimality, and its application hinges on the ability to guess a solution of the HJB equation in a given setting (otherwise, we resort to numerical methods). We next discuss a couple of examples when a solution can be obtained.

5.3.1 The linear quadratic regulator problem

Consider the controlled SDE

$$\mathrm{d}X_t = A(t)X_t\, \mathrm{d}t + B(t)u_t\, \mathrm{d}t + F(t)\, \mathrm{d}W_t, \qquad 0 \leq t \leq 1, \tag{5.3.27}$$

where $A(t)$, $B(t)$, and $F(t)$ are matrices of appropriate shapes. This is a time-varying system, but it is easy to generalize the above discussion to the time-varying case with $f(x,u,t)$ and $g(x,u,t)$ instead of $f(x,u)$ and $g(x,u)$, $q(x,u,t)$ instead of $q(x,u)$, $\mathcal{A}_t^u$ instead of $\mathcal{A}^u$, etc. The HJB equation will then take the form

$$\frac{\partial}{\partial t} V(x,t) + \min_{u \in \mathbb{R}^m} \{\mathcal{A}_t^u V(x,t) + q(x,u,t)\} = 0, \qquad (x,t) \in \mathbb{R}^n \times [0,1] \tag{5.3.28}$$

with the terminal condition $V(x,1) = r(x)$.

We wish to find an admissible control for (5.3.27) to minimize the expected cost

$$J(x,t;\boldsymbol{u}) = \frac{1}{2}\mathbf{E}\left[\int_t^1 \left(X_s^T P(s) X_s + u_s^T Q(s) u_s\right) \mathrm{d}s + X_1^T R X_1 \middle| X_t = x\right] \tag{5.3.29}$$

where the matrices $P(\cdot), R$ are symmetric and positive semidefinite and the matrices $Q(\cdot)$ are symmetric and positive definite. Thus, we have

$$f(x,u,t) = A(t)x + B(t)u,\ g(x,u,t) = F(t); \tag{5.3.30}$$

$$q(x,u,t) = \frac{1}{2}\big(x^T P(t) x + u^T Q(t) u\big),\ r(x) = \frac{1}{2} x^T R x, \tag{5.3.31}$$

which comes with the family of infinitesimal generators

$$\mathcal{A}_t^u \varphi(x,t) = \big(A(t)x + B(t)u\big)^T \nabla\varphi(x,t) + \frac{1}{2}\mathrm{tr}\{F(t)F(t)^T\nabla^2\varphi(x,t)\} \tag{5.3.32}$$

$$= x^T A(t)^T \nabla\varphi(x,t) + \frac{1}{2}\mathrm{tr}\{F(t)F(t)^T\nabla^2\varphi(x,t)\} + u^T B(t)^T\nabla\varphi(x,t). \tag{5.3.33}$$

Now, for any $\xi \in \mathbb{R}^n$,

$$\min_{u\in\mathbb{R}^m}\left\{u^T B(t)^T\xi + \frac{1}{2}u^T Q(t)u\right\} = -\frac{1}{2}\xi^T B(t)Q(t)^{-1}B(t)^T\xi \tag{5.3.34}$$

where the minimum is attained uniquely at $\bar{u}(\xi,t) := -Q(t)^{-1}B(t)^T\xi$. Using this, we can write down the HJB equation:

$$\frac{\partial}{\partial t}V(x,t) = -x^T A(t)^T\nabla V(x,t) + \frac{1}{2}\nabla V(x,t)^T B(t)Q(t)^{-1}B(t)^T\nabla V(x,t)$$
$$- \frac{1}{2}\mathrm{tr}\{F(t)F(t)^T\nabla^2 V(x,t)\} - \frac{1}{2}x^T P(t)x \tag{5.3.35}$$

with the terminal condition $V(x,1) = \frac{1}{2}x^T Rx$. Given the structure of the problem, it is reasonable to guess the quadratic form for $V(x,t)$:

$$V(x,t) = \frac{1}{2}x^T M(t)x + q(t), \tag{5.3.36}$$

where $M(t)$ are symmetric, positive semidefinite matrices and $q(t)$ are nonnegaitve reals. The terminal condition $V(x,1) = \frac{1}{2}x^T Rx$ then leads to $M(1) = R$ and $q(1) = 0$. Substituting

$$\dot{V}(x,t) = \frac{1}{2}x^T\dot{M}(t)x + \dot{q}(t),\ \nabla V(x,t) = M(t)x,\ \nabla^2 V(x,t) = M(t) \tag{5.3.37}$$

into (5.3.35) leads to

$$\frac{1}{2}x^T\dot{M}x + \dot{q} = -\frac{1}{2}x^T\big(A^TM + MA - MBQ^{-1}B^TM + P\big) - \frac{1}{2}\mathrm{tr}\{FF^TM\} \tag{5.3.38}$$

(we have omitted the time variable to avoid notational clutter). This gives us ODEs for $M(t)$ and for $q(t)$:

$$\dot{M} = -A^TM - MA + MBQ^{-1}B^TM - P, \qquad M(1) = R \tag{5.3.39}$$

and

$$\dot{q} = -\frac{1}{2}\mathrm{tr}\{FF^TM\}, \qquad q(1) = 0. \tag{5.3.40}$$

The equation for $M(t)$ is of the familiar Riccati form, while the one for $q(t)$ can be solved in terms of $M(t)$:

$$q(t) = \frac{1}{2}\int_t^1 \mathrm{tr}\{F(s)F(s)^TM(s)\}\,\mathrm{d}s. \tag{5.3.41}$$

Moreover, if we define the matrices $K(t) := Q(t)^{-1}B(t)^T M(t)$, then it is straightforward to show that

$$\mathcal{A}_t^{-K(t)x} V(x,t) + q(x,-K(t)x,t) = \min_{u\in\mathbb{R}^m} \{\mathcal{A}_t^u V(x,t) + q(x,u,t)\}, \tag{5.3.42}$$

so the state feedback control $u_t = -K(t)x$ is optimal. The minimum value of the control cost, for a given initial condition $X_0 = x$, is then equal to

$$V(x,0) = \frac{1}{2}x^T K(0)x + \frac{1}{2}\int_0^1 \operatorname{tr}\{F(t)F(t)^T K(t)\}\,\mathrm{d}t. \tag{5.3.43}$$

5.3.2 Fleming's logarithmic transformation and the Schrödinger bridge problem

Consider the n-dimensional controlled SDE

$$\mathrm{d}X_t = u_t\,\mathrm{d}t + \mathrm{d}W_t, \qquad 0 \le t \le 1. \tag{5.3.44}$$

Given the initial condition $X_0 = x$, we wish to find a control process $\boldsymbol{u} = (u_t)_{0\le t\le 1}$ to minimize the expected cost

$$J(x;\boldsymbol{u}) := \mathbf{E}\left[\frac{1}{2}\int_0^1 |u_t|^2\,\mathrm{d}t + r(X_1)\middle| X_0 = x\right]. \tag{5.3.45}$$

The corresponding cost-to-go is then

$$J(x,t;\boldsymbol{u}) = \mathbf{E}\left[\frac{1}{2}\int_t^1 |u_s|^2\,\mathrm{d}s + r(X_1)\middle| X_t = x\right]. \tag{5.3.46}$$

Here, the effect of the control is to add a drift to a Brownian motion process; the overall control cost consists of the 'control effort' $\frac{1}{2}\int_0^1 |u_t|^2\,\mathrm{d}t$ and a terminal cost $r(X_1)$. Thus, we wish to find a drift that achieves the best trade-off between the expected control effort (how much perturbation is added to the Brownian motion) and the expected terminal cost.

In this set-up, we have $f(x,u) = u$, $g(x,u) = I_n$, $q(x,u) = \frac{1}{2}|u|^2$, and $\mathcal{A}^u\varphi = u^T\nabla\varphi + \frac{1}{2}\Delta\varphi$. Since

$$\min_{u\in\mathbb{R}^n}\left\{u^T\xi + \frac{1}{2}|u|^2\right\} = -\frac{1}{2}|\xi|^2, \tag{5.3.47}$$

the HJB equation takes the form

$$\frac{\partial}{\partial t}V(x,t) + \frac{1}{2}\Delta V(x,t) = \frac{1}{2}|\nabla V(x,t)|^2, \qquad (x,t)\in\mathbb{R}^n\times[0,1] \tag{5.3.48}$$

with the terminal condition $V(x,1) = r(x)$. In order to apply the verification theorem, we need to come up with a solution of (5.3.48), which, on the face of it, looks like a difficult task. However, as first shown by W. Fleming, a simple logarithmic transformation suffices to reduce the problem to

solving a linear second-order PDE. To that end, let us consider the function $h(x,t) := e^{-V(x,t)}$, or $V(x,t) = -\log h(x,t)$. It is straightforward to compute

$$\frac{\partial}{\partial t} h(x,t) = -h(x,t) \frac{\partial}{\partial t} V(x,t), \tag{5.3.49}$$

$$\frac{\partial}{\partial x_i} h(x,t) = -h(x,t) \frac{\partial}{\partial x_i} V(x,t), \tag{5.3.50}$$

$$\begin{aligned} \frac{\partial^2}{\partial x_i^2} h(x,t) &= -\frac{\partial}{\partial x_i}\Big(h(x,t) \frac{\partial}{\partial x_i} V(x,t)\Big) \\ &= -\frac{\partial}{\partial x_i} h(x,t) \frac{\partial}{\partial x_i} V(x,t) - h(x,t) \frac{\partial^2}{\partial x_i^2} V(x,t) \\ &\quad h(x,t)\Big(\Big|\frac{\partial}{\partial x_i} V(x,t)\Big|^2 - \frac{\partial^2}{\partial x_i^2} V(x,t)\Big), \end{aligned} \tag{5.3.51}$$

which gives

$$\nabla h(x,t) = -h(x,t) \nabla v(x,t), \qquad \Delta h(x,t) = h(x,t)\big(|\nabla V(x,t)|^2 - \Delta V(x,t)\big). \tag{5.3.52}$$

Using these relations, we can obtain a PDE for h:

$$\frac{\partial}{\partial t} h(x,t) = -h(x,t) \frac{\partial}{\partial t} V(x,t) \tag{5.3.53}$$

$$= \frac{1}{2} h(x,t)\big(\Delta V(x,t) - |\nabla V(x,t)|^2\big) \tag{5.3.54}$$

$$= -\frac{1}{2} \Delta h(x,t), \tag{5.3.55}$$

with the terminal condition $h(x,1) = e^{-r(x)}$. It is convenient to reverse the time by defining $\tilde{h}(x,t) := h(x,1-t)$, so that $\tilde{h}(x,0) = e^{-r(x)}$ and

$$\frac{\partial}{\partial t} \tilde{h}(x,t) + \frac{1}{2} \Delta \tilde{h}(x,t) = 0. \tag{5.3.56}$$

This can be solved immediately using the Feynman–Kac formula:

$$\tilde{h}(x,t) = \mathbf{E}[e^{-r(x+W_t)}], \tag{5.3.57}$$

where $(W_t)_{0 \le t \le 1}$ is a standard n-dimensional Brownian motion. Since $W_t \sim \mathcal{N}(0, tI_n)$, we can express this as a Gaussian integral: If $Z \sim \mathcal{N}(0, I_n)$, then

$$\tilde{h}(x,t) = \mathbf{E}[e^{-r(x+\sqrt{t}Z)}] = \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} e^{-r(x+\sqrt{t}z)} e^{-|z|^2} \, \mathrm{d}z. \tag{5.3.58}$$

Consequently,

$$V(x,t) = -\log h(x,t) \tag{5.3.59}$$

$$= -\log \tilde{h}(x,1-t) \tag{5.3.60}$$

$$= -\log \mathbf{E}[e^{-r(x+\sqrt{1-t}Z)}], \tag{5.3.61}$$

and the state feedback control $u_t = -\nabla V(x,t)$ is optimal. The minimum expected cost is given by

$$V(x,0) = -\log \mathbf{E}[e^{-r(x+Z)}]. \tag{5.3.62}$$

An interesting application of this control problem is as follows: Let a strictly positive probability density p on $\mathbb{R}^n$ be given. Consider all adapted control processes $\boldsymbol{u} = (u_t)_{0\le t\le 1}$, such that p is the probability density of

$$X_1 = \int_0^1 u_t \,\mathrm{d}t + W_1, \tag{5.3.63}$$

and among these find the control that minimizes the expected effort $\mathbf{E}[\frac{1}{2}\int_0^1 |u_t|^2 \,\mathrm{d}t]$. This is an instance of the so-called *Schrödinger bridge problem*, where the goal is to steer the stochastic evolution of a system of particles between $t = 0$ and $t = 1$ to a given distribution at $t = 1$, while making sure that the trajectories of the particles are as close to Brownian as possible. To see how this problem connects to the above optimal control problem, let $\gamma(x) = \frac{1}{(2\pi)^{n/2}} e^{-|x|^2/2}$ denote the standard Gaussian density on $\mathbb{R}^n$ and consider the function

$$r(x) := -\log \frac{p(x)}{\gamma(x)} = -\log p(x) + \frac{1}{2}|x|^2 + \frac{n}{2}\log(2\pi). \tag{5.3.64}$$

(since $p > 0$ everywhere, r is well-defined). Then $\mathbf{E}[e^{-r(Z)}] = 1$ for $Z \sim \mathcal{N}(0, I_n)$, and therefore

$$\min_{\boldsymbol{u}} \mathbf{E}\left[\frac{1}{2}\int_0^1 |u_t|^2 \,\mathrm{d}t + r(X_1)\right] = -\log \mathbf{E}[e^{-r(Z)}] = 0. \tag{5.3.65}$$

Now, if $\boldsymbol{u} = (u_t)_{0\le t\le 1}$ is any control that ensures $X_1 \sim p$, then

$$\mathbf{E}\left[\frac{1}{2}\int_0^1 |u_t|^2 \,\mathrm{d}t + r(X_1)\right] = \mathbf{E}\left[\frac{1}{2}\int_0^1 |u_t|^2 \,\mathrm{d}t\right] + \int_{\mathbb{R}^n} r(x)p(x)\,\mathrm{d}x \tag{5.3.66}$$

$$= \mathbf{E}\left[\frac{1}{2}\int_0^1 |u_t|^2 \,\mathrm{d}t\right] - \int_{\mathbb{R}^n} p(x)\log\frac{p(x)}{\gamma(x)}\,\mathrm{d}x \tag{5.3.67}$$

$$= \mathbf{E}\left[\frac{1}{2}\int_0^1 |u_t|^2 \,\mathrm{d}t\right] - D(p\|\gamma), \tag{5.3.68}$$

where $D(p\|\gamma)$ is the *relative entropy* (or *Kullback–Leibler divergence*) between the densities p and γ, so we get the following lower bound on the required control effort:

$$\boldsymbol{u} \text{ gives } X_1 \sim p \qquad \Rightarrow \qquad \mathbf{E}\left[\frac{1}{2}\int_0^1 |u_t|^2 \,\mathrm{d}t\right] \ge D(p\|\gamma). \tag{5.3.69}$$

Moreover, the control that achieves the minimum in (5.3.65) also solves our optimum steering problem, i.e., attains equality in (5.3.69); this will be shown in a homework problem.

5.4 Optimal control with partial observations

We now consider the setting where we only have noisy observations of the state in (5.3.1), i.e., we have the system

$$\mathrm{d}X_t = f(X_t, u_t)\,\mathrm{d}t + g(X_t, u_t)\,\mathrm{d}W_t \tag{5.4.1a}$$
$$\mathrm{d}Y_t = h(X_t, t)\,\mathrm{d}t + \mathrm{d}V_t, \tag{5.4.1b}$$

where (W_t) and (V_t) are two independent Brownian motion processes. We still face the problem of minimizing the expected cost

$$J(\boldsymbol{u}) = \mathbf{E}\left[\int_0^1 q(X_t, u_t)\,\mathrm{d}t + r(X_1)\right], \tag{5.4.2}$$

where the initial condition X_0 is now random. But now the control $\boldsymbol{u}$ has to be adapted to the filtration $(\mathcal{F}_t^Y)$ generated by the observation process (Y_t). In fact, we have to be a bit more careful here, since the choice of the control influences the state process and hence the observation process, which means that the relevant filtration will generally depend on $\boldsymbol{u}$ as well. In other words, we should be more precise and indicate the dependence on the control everywhere, as in

$$\mathrm{d}X_t^{\boldsymbol{u}} = f(X_t^{\boldsymbol{u}}, u_t)\,\mathrm{d}t + g(X_t^{\boldsymbol{u}}, u_t)\,\mathrm{d}W_t \tag{5.4.3a}$$
$$\mathrm{d}Y_t^{\boldsymbol{u}} = h(X_t^{\boldsymbol{u}}, u_t)\,\mathrm{d}t + \mathrm{d}V_t, \tag{5.4.3b}$$

and the control $\boldsymbol{u}$ is constrained to be adapted to the filtration $(\mathcal{F}_t^{Y,\boldsymbol{u}})$ generated by $(Y_t^{\boldsymbol{u}})$.

The obvious step here is to make use of the nonlinear filter $\pi_t^{\boldsymbol{u}}(\cdot)$, where for any function $\psi(x,u)$ we have

$$\pi_t^{\boldsymbol{u}}[\psi(\cdot, u_t)] = \mathbf{E}[\psi(X_t^{\boldsymbol{u}}, u_t)|\mathcal{F}_t^{Y,\boldsymbol{u}}]. \tag{5.4.4}$$

Using this definition and the law of iterated expectations, we can express the expected cost $J(\boldsymbol{u})$ as

$$J(\boldsymbol{u}) = \mathbf{E}\left[\int_0^1 \pi_t^{\boldsymbol{u}}(q(\cdot, u_t))\,\mathrm{d}t + \pi_1^{\boldsymbol{u}}(r)\right], \tag{5.4.5}$$

where $\boldsymbol{u}$ is any $(\mathcal{F}_t^{Y,\boldsymbol{u}})$-adapted control. This reframes the problem of controlling the hidden state $X_t^{\boldsymbol{u}}$ on the basis of observations $Y_t^{\boldsymbol{u}}$ into one with the fully observed state given by $\pi_t^{\boldsymbol{u}}$, and the control can now be based on the filter output. This idea is known as the *separation principle*, where the problem of choosing the control is *separated* from the problem of estimating the hidden state by filtering the noisy observations. However, any gain due to this reframing is illusory, since now we are confronted with the problem of computing the nonlinear filter, which is already quite difficult (and generally infinite-dimensional). Thus, in general, one has to resort to numerical approximations. The case of linear SDEs for X_t and Y_t, quadratic costs, and Gaussian initial condition is a pleasant exception, although even there we have to be careful.

5.4.1 The linear-quadratic-Gaussian (LQG) control problem

Let us consider the following special case of (5.4.1):

$$\mathrm{d}X_t = A(t)X_t\,\mathrm{d}t + B(t)u_t\,\mathrm{d}t + F(t)\,\mathrm{d}W_t \tag{5.4.6a}$$
$$\mathrm{d}Y_t = H(t)X_t\,\mathrm{d}t + G(t)\,\mathrm{d}V_t, \tag{5.4.6b}$$

with Gaussian initial condition $X_0 \sim \mathcal{N}(m_0, K_0)$. Here, as before, (W_t) and (V_t) are two independent Brownian motion processes of appropriate dimensions that are also independent of X_0. Also, we assume that the matrices $G(t)G(t)^T$ are positive definite for all t, just like in the Kalman–Bucy filter setting. We are not indicating the dependence on $\boldsymbol{u}$ explicitly, as we did in (5.4.3) to keep the notation clean, but it should be kept in mind. The goal is to minimize the expected cost

$$J(\boldsymbol{u}) = \frac{1}{2}\mathbf{E}\left[\int_0^1 \big(X_t^T P(t) X_t \,\mathrm{d}t + u_t^T Q(t) u_t\big)\,\mathrm{d}t + X_1^T R X_1\right], \tag{5.4.7}$$

where the matrices $P(t), Q(t), R$ satisfy the same conditions as in Section 5.3.1, but now the control $\boldsymbol{u}$ must be adapted to $\mathcal{F}_t^{Y,\boldsymbol{u}}$. We will derive the solution shortly, but, in a nutshell, the idea is to obtain an estimate $\hat{X}_t$ of the state X_t using the Kalman filter, and then solve the fully observed LQR problem for controlling $\hat{X}_t$, with the solution given in the estimated state feedback form $u_t = -K(t)\hat{X}_t$. This is, of course, a manifestation of the separation principle: The task of estimating the state and the task of choosing the control are separated.

To begin the analysis, let us first consider the case of zero controls (obtained by setting $u_t \equiv 0$ in (5.4.6a)) which is described by the state equation

$$\mathrm{d}X_t^0 = A(t)X_t^0\,\mathrm{d}t + F(t)\,\mathrm{d}W_t \tag{5.4.8}$$

and the observation equation

$$\mathrm{d}Y_t^0 = H(t)X_t^0\,\mathrm{d}t + G(t)\,\mathrm{d}V_t \tag{5.4.9}$$

with the initial conditions $X_0^0 = X_0$, $Y_0^0 = 0$. Let $\mathcal{F}_t^{Y,0}$ denote the σ-algebra generated by $(Y_s^0)_{0\le s\le t}$. Then we can write down the Kalman–Bucy filter equation for $\hat{X}_t^0 = \mathbf{E}[X_t^0|\mathcal{F}_t^{Y,0}]$:

$$\mathrm{d}\hat{X}_t^0 = A(t)\hat{X}_t^0\,\mathrm{d}t + L(t)\big(\mathrm{d}Y_t^0 - H(t)\hat{X}_t^0\,\mathrm{d}t\big) \tag{5.4.10}$$

with the initial condition $\hat{X}_0^0 = m_0$, and let $\Pi(t)$ be the error covariance $\mathbf{E}[(\hat{X}_t^0 - X_t^0)(\hat{X}_t^0 - X_t^0)^T]$. If we now define the processes X_t^1, Y_t^1 by

$$\mathrm{d}X_t^1 = \big(A(t)X_t^1 + B(t)u_t\big)\,\mathrm{d}t, \tag{5.4.11}$$

$$\mathrm{d}Y_t^1 = H(t)X_t^1\,\mathrm{d}t \tag{5.4.12}$$

with $X_0^1 = 0$ and $Y_0^1 = 0$, then evidently

$$X_t = X_t^0 + X_t^1, \qquad Y_t = Y_t^0 + Y_t^1. \tag{5.4.13}$$

We then have the following result:

Lemma 4. *Suppose that the control $\boldsymbol{u}$ is adapted to $\mathcal{F}_t^{Y,0}$, and the σ-algebras $\mathcal{F}_t^{Y,0}$ and $\mathcal{F}_t^{Y,\boldsymbol{u}}$ are equal for all t. Then:*

1. *The conditional distribution of X_t given $\mathcal{F}_t^{Y,\boldsymbol{u}}$ is Gaussian, with conditional mean $\hat{X}_t = \mathbf{E}[X_t|\mathcal{F}_t^{Y,\boldsymbol{u}}] = \hat{X}_t^0 + X_t^1$ and conditional covariance $\Pi(t)$.*
2. *The conditional mean $\hat{X}_t$ satisfies the SDE*

$$\mathrm{d}\hat{X}_t = \big(A(t)\hat{X}_t + B(t)u_t\big)\,\mathrm{d}t + L(t)\big(\mathrm{d}Y_t^0 - H(t)\hat{X}_t^0\,\mathrm{d}t\big). \tag{5.4.14}$$

Proof. The proof is straightforward: Since the σ-algebras $\mathcal{F}_t^{Y,0}$ and $\mathcal{F}_t^{Y,\boldsymbol{u}}$ are equal, conditional expectations given $\mathcal{F}_t^{Y,\boldsymbol{u}}$ and $\mathcal{F}_t^{Y,0}$ are equal. By the same token, since the control u_t is adapted to $\mathcal{F}_t^{Y,0}$, so are X_t^1 and Y_t^1. Everything then follows from the analysis of the Kalman–Bucy filter. $\square$

Let $\boldsymbol{u}$ satisfy the conditions of the Lemma. We can rewrite the expected cost of $\boldsymbol{u}$ in terms of $\hat{X}_t$ and the error $\tilde{X}_t = X_t - \hat{X}_t$. For the latter, we have

$$X_t - \hat{X}_t = X_t^0 + X_t^1 - (\hat{X}_t^0 + X_t^1) = X_t^0 - \hat{X}_t^0, \tag{5.4.15}$$

so $\tilde{X}_t$ is zero-mean and independent of $\hat{X}_t$ and u_t. Then

$$J(\boldsymbol{u}) = \underbrace{\mathbf{E}\left[\int_0^1 \left(\hat{X}_t^T P(t)\hat{X}_t + u_t^T Q(t) u_t\right) \mathrm{d}t + \hat{X}_1^T R \hat{X}_1\right]}_{=:J'(\boldsymbol{u})} + \mathbf{E}\left[\int_0^1 \tilde{X}_t^T P(t)\tilde{X}_t \,\mathrm{d}t + \tilde{X}_1^T R \tilde{X}_1\right] \tag{5.4.16}$$

$$= \hat{J}(\boldsymbol{u}) + \int_0^1 \mathrm{tr}\{\Pi(t)P(t)\}\,\mathrm{d}t + \mathrm{tr}\{\Pi(1)R\}, \tag{5.4.17}$$

where the last two terms are nonrandom and determined by the error covariance matrices $\Pi(t)$. The modified cost $J'(\boldsymbol{u})$ involves the Kalman–Bucy filter output $\hat{X}_t$ and the control u_t, and this is an instance of the linear quadratic regulator since the forward differential of the innovations process $\nu_t^0 := Y_t^0 - \int_0^t H(s)\hat{X}_s^0 \,\mathrm{d}s$ can be written as

$$\mathrm{d}\nu_t^0 = \sqrt{(G(t)G(t)^T)^{-1}}\,\mathrm{d}\hat{W}_t^0 \tag{5.4.18}$$

for some Brownian motion process $\hat{W}_t^0$ adapted to $\mathcal{F}_t^{Y,0}$.

5.5 Problems

1. Consider a linear SDE

$$\mathrm{d}X_t = A(t)X_t\,\mathrm{d}t + G(t)\,\mathrm{d}W_t \tag{5.5.1}$$

for an n-dimensional diffusion process $(X_t)_{t\geq 0}$ driven by an m-dimensional Brownian motion $(W_t)_{t\geq 0}$, where $A(t) \in \mathbb{R}^{n\times n}$ and $G(t) \in \mathbb{R}^{n\times m}$ are time-varying matrices. Show that, for any $0 \leq s \leq t$,

$$X_t = \Phi(t,s)X_s + \int_s^t \Phi(t,r)G(r)\,\mathrm{d}W_r, \tag{5.5.2}$$

where $\Phi(t,s)$ is the fundamental matrix of the linear time-varying system $\dot{x}(t) = A(t)x(t)$, i.e., it solves the matrix ODE

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi(t,s) = A(t)\Phi(t,s), \qquad t \geq s \tag{5.5.3}$$

with the initial condition $\Phi(s,s) = I_n$.

2. Let $\tilde{X}_t$ and ν_t be the state estimation error and the innovation processes of the Kalman–Bucy filter, cf. Section 5.1. Prove that $\tilde{X}_t$ is orthogonal to $(\nu_s)_{0\le s\le t}$, i.e.,

$$\mathbf{E}[\tilde{X}_t \nu_s^T] = 0, \qquad 0 \le s \le t. \tag{5.5.4}$$

Hint: Use (5.1.15) and Problem 1.

3. Consider the Schrödinger bridge problem discussed in Section 5.3.2. For $t > 0$, let $\gamma_t(x) := \frac{1}{(2\pi t)^{n/2}} e^{-|x|^2/2t}$.

(i) For the SDE

$$\mathrm{d}X_t = -\nabla V(X_t, t)\,\mathrm{d}t + \mathrm{d}W_t, \tag{5.5.5}$$

where $V(x,t)$ is the solution of the HJB equation (5.3.48), show that the density $p_t(x) = (1/C_t)\gamma_t(x)e^{-V(x,t)}$, where C_t are the normalization constants, solves the forward Kolmogorov equation

$$\frac{\partial}{\partial t} p_t(x) = \nabla \cdot \big(\nabla V(x,t) p_t(x)\big) + \frac{1}{2}\Delta p_t(x), \qquad 0 \le t \le 1 \tag{5.5.6}$$

with the initial condition $p_0(x) = \delta(x)$.

Note: that normalization C_t constant is important!

(ii) If the terminal cost $r(x)$ is equal to $-\log \frac{p(x)}{\gamma(x)}$, where $\gamma(x) = \gamma_1(x)$ is the standard Gaussian density on $\mathbb{R}^n$, show that X_1 (with initial condition $X_0 = 0$) has density p and thus show that the feedback control $u_t = -\nabla V(X_t, t)$ attains equality in (5.3.69).

Chapter 6

Optimization

Function optimization is a basic task in a variety of engineering applications and settings. A general optimization problem can be formulated as follows:

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } x \in C \end{aligned}$$

where f is a real-valued function defined on some set $\mathcal{X}$ and where $C \subseteq \mathcal{X}$ is a constraint set. Some examples are:

1. Minimizing a linear function $f(x) = c^T x$ of $x \in \mathbb{R}^n$, where $c \in \mathbb{R}^n$ is a fixed vector, subject to linear inequality constraints of the form

$$a_i^T x \leq b_i, \qquad 1 \leq i \leq m. \tag{6.0.1}$$

   This is the classical problem of linear programming, common in such settings as resource allocation.

2. Statistical estimation problems involve minimizing quadratic functions of the form $f(x) = x^T P x + c^T x$ subject to various constraints, e.g., linear inequality constraints as in (6.0.1).

3. Stochastic optimization problems, where $f(x)$ has the form of an expectation w.r.t. some random variable Z, i.e., $f(x) = \mathbf{E}[F(x, Z)]$. For example, if ξ is a random couple (ξ, Y), where ξ is a vector of features and Y is a vector of responses, the goal is to predict Y on the basis of ξ, and we have a family of candidate predictors $\hat{Y} = h(\xi, x)$, then we could take

$$F(x, Z) = F(x, (\xi, Y)) = |Y - h(\xi, x)|^2.$$

4. Optimal control problems of the sort considered in Chapter 5 are infinite-dimensional optimization problems: The set $\mathcal{X}$ is (a subset of) some function space.

5. Other infinite-dimensional optimization problems, e.g., when $\mathcal{X}$ is a set of probability measures on some sample space Ω, such as the path space $\Omega = C([0, T]; \mathbb{R}^d)$.

In this chapter, we will explore the use of SDEs in the context of certain types of optimization problems.

6.1 Langevin dynamics

Consider the problem of globally minimizing a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, i.e., the goal is to find a point $\bar{x} \in \mathbb{R}^n$ such that $f(\bar{x}) \leq f(x)$ for all $x \in \mathbb{R}^n$. Assuming that at least one such global minimizer $\bar{x}$ exists (at the very least, we would assume that f is bounded from below), it would have to satisfy $\nabla f(\bar{x}) = 0$. (Note that this is only a necessary condition for (local) optimality; if f is twice differentiable, then $\bar{x}$ is a (local) minimizer if and only if $\nabla f(\bar{x}) = 0$ and the Hessian $\nabla^2 f(\bar{x})$ is positive semidefinite.) This observation underlies the use of gradient descent methods: Pick some initial point $x_0 \in \mathbb{R}^n$ and generate the iterates

$$x_{k+1} = x_k - \eta_k \nabla f(x_k), \qquad k = 0, 1, \ldots \tag{6.1.1}$$

where $(\eta_k)_{k\geq 0}$ is a decreasing sequence of positive *step sizes*. If the step sizes are sufficiently small, then we can consider a continuous-time idelization of (6.1.1), the *gradient flow*

$$\dot{x}(t) = -\nabla f(x(t)), \qquad t \geq 0 \tag{6.1.2}$$

A simple calculation shows that the value of $f(x(t))$ decreases along the trajectory of (6.1.2):

$$\frac{\mathrm{d}}{\mathrm{d}t} f(x(t)) = \nabla f(x(t))^T \dot{x}(t) \tag{6.1.3}$$

$$= -|\nabla f(x(t))|^2 \tag{6.1.4}$$

$$\leq 0. \tag{6.1.5}$$

However, unless the function f has additional properties, such as convexity, there can be no guarantee that $x(t)$ will converge, let alone converge to a global minimizer. In fact, depending on the initial condition $x(0)$, the trajectory $x(t)$ may get trapped in the neighborhood of some local minimizer.

Intuitively, the reason for this behavior is that gradient descent methods are, by their nature, local: At each time t, the negative gradient $-\nabla f(x(t))$ points in the direction of greatest initial decrease of f (thus the name "steepest descent" that is often attached to gradient methods), but only the points $x(t)$ in a small vicinity of the trajectory of (6.1.2) can be explored in this way. A way to get around this is to inject some stochasticity into the dynamics and consider, instead of (6.1.2), the SDE

$$\mathrm{d}X_t = -\nabla f(X_t)\,\mathrm{d}t + \sqrt{2\varepsilon}\,\mathrm{d}W_t, \tag{6.1.6}$$

where $\varepsilon > 0$ is a small parameter, typically referred to as the temperature. The SDE (6.1.6) is called the (overdamped) *Langevin dynamics* with *potential* f. The name comes from the work of Paul Langevin on Brownian motion, where the quadratic potential $f(x) = c|x|^2$ was used to model the frictional forces in the surrounding medium. To provide some motivation for the form of (6.1.6), consider the following *stochastic relaxation* of the problem of minimizing f: Instead of searching for a global minimizer $\bar{x}$, let us sample a *random* point $\bar{X}$ from the density

$$\pi^\varepsilon(x) := \frac{1}{Z^\varepsilon} \exp\Big(-\frac{1}{\varepsilon} f(x)\Big), \tag{6.1.7}$$

where $Z^\varepsilon := \int_{\mathbb{R}^n} e^{-(1/\varepsilon) f(x)}\,\mathrm{d}x$ is the normalization constant (we assume that $f(x)$ is sufficiently well-behaved as $|x| \to \infty$ for Z^ε to be finite). This explains why we have used the term 'temperature' to refer to the parameter ε: In statistical physics, probability densities of the form (6.1.7) are referred

to as *Gibbs densities* and are used to model the behavior of large systems in thermal equilibrium, where $f(x)$ is the 'energy' of a 'system configuration' $x \in \mathbb{R}^n$ and ε is the absolute temperature. Under some regularity conditions on f, it can be shown that

$$\mathbf{E}[f(\bar{X})] \leq \min_x f(x) + n\varepsilon \log \frac{C}{n\varepsilon} \tag{6.1.8}$$

for some constnat $C > 0$ that depends on f. Moreover, π^ε is an equilibrium density of (6.1.6), i.e., if $X_0 \sim \pi^\varepsilon$, then $X_t \sim \pi^\varepsilon$ for all t. To see this, let us look at the Fokker–Planck equation for the density $\rho_t(x)$ of X_t when X_0 is sampled from some density ρ_0:

$$\frac{\partial}{\partial t}\rho_t(x) = \nabla \cdot \big(\nabla f(x)\rho_t(x)\big) + \varepsilon \Delta \rho_t(x). \tag{6.1.9}$$

Since $\nabla \log \pi^\varepsilon(x) = -\frac{1}{\varepsilon}\nabla f(x)$, it follows that π^ε is a stationary solution of (6.1.9). Moreover, under certain regularity conditions on f, it is possible to show that the probability density of X_t converges to π^ε as $t \to \infty$ for any initial density ρ_0 of X_0. While the precise argument requires a great deal of functional analysis, we can give a brief sketch illustrating the basic ideas.

6.1.1 Convergence to equilibrium and the spectral gap

Let $p_t(x_0, x)$ denote the transition density of (6.1.6), i.e.,

$$\mathbf{E}[\varphi(X_t)|X_0 = x_0] = \int_{\mathbb{R}^n} \varphi(x) p_t(x_0, x)\,\mathrm{d}x \tag{6.1.10}$$

for all smooth functions $\varphi : \mathbb{R}^n \to \mathbb{R}$. It satisfies the Fokker–Planck equation

$$\frac{\partial}{\partial t}p_t(x_0, x) = \nabla \cdot \big(\nabla f(x) p_t(x_0, x)\big) + \varepsilon \Delta p_t(x_0, 0) \tag{6.1.11}$$

with the initial condition $\lim_{t\to 0} p_t(x_0, x) = \delta(x - x_0)$. (As usual, the operators ∇ and Δ act on the x variable, while x_0 remains fixed.) Then, for sufficiently well-behaved f, it can be shown that we can express $p_t(x_0, x)$ as

$$p_t(x_0, x) = \pi^\varepsilon(x) \sum_{i=0}^{\infty} e^{-\lambda_i t} \psi_i(x)\psi_i(x_0), \tag{6.1.12}$$

where $\lambda_i \geq 0$ and ψ_i are the eigenvalues and the normalized eigenfunctions of the *Sturm–Liouville equation*

$$\varepsilon \nabla \cdot \big(\pi^\varepsilon(x)\nabla\psi(x)\big) + \lambda \pi^\varepsilon(x)\psi(x) = 0, \tag{6.1.13}$$

i.e., both the function ψ and the constant λ must be solved for, and, since (6.1.13) is linear in ψ, we require that

$$\int_{\mathbb{R}^n} \pi^\varepsilon(x)|\psi(x)|^2 = 1. \tag{6.1.14}$$

The basic idea is to attempt to solve (6.1.11) using separation of variables. To that end, consider an ansatz of the form $h(x, t) = g(t)\pi^\varepsilon(x)\psi(x)$. Then

$$\frac{\partial}{\partial t}h(x, t) = \pi^\varepsilon(x)\psi(x)\frac{\mathrm{d}}{\mathrm{d}t}g(t), \tag{6.1.15}$$

while

$$\nabla f(x)\cdot h(x,t)+\varepsilon\nabla h(x,t)=g(t)\left(\psi(x)\Big(\underbrace{\pi^{\varepsilon}(x)\nabla f(x)+\varepsilon\nabla\pi^{\varepsilon}(x)}_{=0}\Big)+\varepsilon\pi^{\varepsilon}(x)\nabla\psi(x)\right) \tag{6.1.16}$$

$$=\varepsilon g(t)\pi^{\varepsilon}(x)\nabla\psi(x). \tag{6.1.17}$$

Thus, if $h(x,t)$ is a solution of (6.1.11), then it has to satisfy

$$\frac{1}{g(t)}\frac{\mathrm{d}}{\mathrm{d}t}g(t)=\frac{1}{\pi^{\varepsilon}(x)\psi(x)}\varepsilon\nabla\cdot\big(\pi^{\varepsilon}(x)\nabla\psi(x)\big) \tag{6.1.18}$$

The left-hand side of (6.1.18) is a function of t only, while the right-hand side is a function of x only. Thus, they are both equal to some constant $-\lambda$, so we can take $g(t)=e^{-\lambda t}$, which implies that ψ and λ must satisfy (6.1.13).

Problems of this type are studied in functional analysis, and in particular in the theory of *Schrödinger operators*, named this way because they appear in the context of Schrödinger's equation of quantum mechanics. Using the theory of Schrödinger operators, it is possible to show that there are only *countably many* solutions of (6.1.13), i.e., the eigenfunction-eigenvalue pairs (ψ_i,λ_i). It is immediate that $\psi_0(x)\equiv 1$ is an eigenfunction with $\lambda_0=0$, and it can be shown that all other eigenvalues $\lambda_1\le\lambda_2\le\dots$ are positive. The first nonzero eigenvalue λ_1 is important because it controls the rate at which $p_t(x_0,x)$ converges to $\pi^{\varepsilon}(x)$ as $t\to\infty$. To see this, let us write

$$|p_t(x_0,x)-\pi^{\varepsilon}(x)|=\left|\sum_{i=1}^{\infty}e^{-\lambda_i t}\psi_i(x)\psi_i(x_0)\right| \tag{6.1.19}$$

$$=e^{-\lambda_1 t}\left|\sum_{i=1}^{\infty}e^{-(\lambda_i-\lambda_1)t}\psi_i(x)\psi_i(x_0)\right| \tag{6.1.20}$$

Now,

$$|p_1(x,x)-\pi^{\varepsilon}(x)|=\left|\sum_{i=1}^{\infty}e^{-\lambda_i}\psi_i^2(x)\right|, \tag{6.1.21}$$

so if we define

$$k(x):=e^{\lambda_1}|p_1(x,x)-\pi^{\varepsilon}(x)| \tag{6.1.22}$$

$$=\sum_{i=1}^{\infty}e^{-(\lambda_i-\lambda_1)}\psi_i^2(x) \tag{6.1.23}$$

then, for $t>1$,

$$\left|\sum_{i=1}^{\infty}e^{-(\lambda_i-\lambda_1)t}\psi_i(x)\psi_i(x_0)\right|\le\sqrt{\sum_{i=1}^{\infty}e^{-(\lambda_i-\lambda_1)t}\psi_i^2(x)\sum_{j}^{\infty}e^{-(\lambda_j-\lambda_1)t}\psi_j^2(x_0)} \tag{6.1.24}$$

$$\le\sqrt{k(x)k(x_0)} \tag{6.1.25}$$

$$\le\sup_{x\in\mathbb{R}^n}k(x), \tag{6.1.26}$$

where the first inequality uses Cauchy–Schwarz. Thus, for $t > 1$,

$$|p_t(x_0, x) - \pi^\varepsilon(x)| \le K e^{-\lambda_1 t}, \qquad K := \sup_{x \in \mathbb{R}^n} k(x) \tag{6.1.27}$$

where we assume that f is such that the constant K defined above is finite. Then, if ρ_0 is the initial density of X_0, we have

$$\begin{aligned} |\rho_t(x) - \pi^\varepsilon(x)| &= \left| \int_{\mathbb{R}^n} p_t(x_0, x) \rho_0(x_0) \, \mathrm{d}x_0 - \pi^\varepsilon(x) \right| && (6.1.28) \\ &\le \int_{\mathbb{R}^n} \rho_0(x_0) |p_t(x_0, x) - \pi^\varepsilon(x)| \, \mathrm{d}x_0 && (6.1.29) \\ &\le K e^{-\lambda_1 t}. && (6.1.30) \end{aligned}$$

The constant λ_1 can be characterized as follows. Let $\mathcal{S}$ be the space of all C^1 functions $\varphi : \mathbb{R}^n \to \mathbb{R}$, such that

$$\int_{\mathbb{R}^n} \pi^\varepsilon(x) \varphi(x) \, \mathrm{d}x = 0 \qquad \text{and} \qquad \int_{\mathbb{R}^n} \pi^\varepsilon(x) |\varphi(x)|^2 \, \mathrm{d}x < \infty. \tag{6.1.31}$$

It is easy to see that $\mathcal{S}$ is a vector space, and all the eigenfunctions $\psi_1, \psi_2, \dots$ are the elements of $\mathcal{S}$. Now, if we consider (6.1.13) for (ψ_i, λ_i), $i \ge 1$, multiply both sides by ψ_i and integrate, we get

$$\begin{aligned} \lambda_i &= -\varepsilon \int_{\mathbb{R}^n} \psi_i(x) \nabla \cdot \big(\pi^\varepsilon(x) \nabla \psi_i(x) \big) \, \mathrm{d}x && (6.1.32) \\ &= \varepsilon \int_{\mathbb{R}^n} \pi^\varepsilon(x) |\nabla \psi_i(x)|^2 \, \mathrm{d}x, && (6.1.33) \end{aligned}$$

where the second step follows after integrating by parts (and assuming that ψ_i, π^ε decay sufficiently rapidly as $|x| \to \infty$). Thus, for any $\varphi \in \mathcal{S}$,

$$\lambda_1 \le \varepsilon \frac{\int_{\mathbb{R}^n} \pi^\varepsilon(x) |\nabla \varphi(x)|^2 \, \mathrm{d}x}{\int_{\mathbb{R}^n} \pi^\varepsilon(x) |\varphi(x)|^2 \, \mathrm{d}x}, \tag{6.1.34}$$

but, since equality holds for $\varphi = \psi_1$, we obtain

$$\lambda_1 = \varepsilon \min_{\varphi \in \mathcal{S}} \frac{\int_{\mathbb{R}^n} \pi^\varepsilon(x) |\nabla \varphi(x)|^2 \, \mathrm{d}x}{\int_{\mathbb{R}^n} \pi^\varepsilon(x) |\varphi(x)|^2 \, \mathrm{d}x}. \tag{6.1.35}$$

The constant λ_1 is known as the *Poincaré constant* (or *spectral gap*) associated to the Langevin diffusion (6.1.6), and there exist many methods for obtaining upper and lower bounds for it. Generally, it will scale rather poorly with problem dimension n and inverse temperature $1/\varepsilon$—e.g., if f has multiple minima as well as saddle points (i.e., the points x such that $\nabla f(x) = 0$ but the Hessian $\nabla^2 f(x)$ has both positive and negative eigenvalues), then the spectral gap will scale like e^{-n} in n and like $e^{-1/\varepsilon}$ in ε. When f has enough 'curvature,' however (in particular, if it has no local minima, and thus all minima are global), much better dependence on n and $1/\varepsilon$ can be obtained.

6.1.2 The relative entropy and the logarithmic Sobolev inequality

Another way to quantify the convergence of ρ_t to π^ε is through an information-theoretic quantity known as the *relative entropy* or the *Kullback–Leibler divergence*. For our purposes, the following definition can be given: Let ρ be a probability density on $\mathbb{R}^n$. Then the relative entropy between ρ and π^ε is given by

$$D(\rho\|\pi^\varepsilon) := \int_{\mathbb{R}^n} \rho(x) \log \frac{\rho(x)}{\pi^\varepsilon(x)} \,\mathrm{d}x. \tag{6.1.36}$$

Detailed discussions of this quantity can be found in many texts on information theory; here we note the important property that $D(\rho\|\pi^\varepsilon)$ is always nonnegative, and it is equal to zero if and only if $\rho = \pi^\varepsilon$. Another important quantity is the *relative Fisher information* between ρ and π^ε, defined as

$$I(\rho\|\pi^\varepsilon) := \int_{\mathbb{R}^n} \rho(x) \Big| \nabla \log \frac{\rho(x)}{\pi^\varepsilon(x)} \Big|^2 \,\mathrm{d}x, \tag{6.1.37}$$

which is evidently nonnegative and vanishes if and only if $\nabla \log \frac{\rho}{\pi^\varepsilon} \equiv 0$.

Let us now consider the Langevin diffusion (6.1.6) with X_0 having a density ρ_0. The evolution of the densities ρ_t is governed by the Fokker–Planck equation (6.1.9), and it can be shown that ρ_t is positive everywhere if ρ_0 is. It will be convenient to express the Fokker–Planck equation in the following form:

$$\frac{\partial}{\partial t}\rho_t(x) = -\nabla \cdot j_t(x), \tag{6.1.38}$$

where

$$j_t(x) := -\rho_t(x)\nabla f(x) - \varepsilon \nabla \rho_t(x) \tag{6.1.39}$$

is referred to as the *probability flux* or *current*. In equilibrium, i.e., when $\rho_t = \pi^\varepsilon$ for all t, the probability flux vanishes. Hence, writing $\rho_t = \pi^\varepsilon \cdot \frac{\rho_t}{\pi^\varepsilon}$, we have

$$j_t(x) = -\frac{\rho_t(x)}{\pi^\varepsilon(x)} \Big(\underbrace{\pi^\varepsilon(x)\nabla f(x) + \varepsilon \nabla \pi^\varepsilon(x)}_{=0} \Big) - \varepsilon \pi^\varepsilon(x) \nabla \frac{\rho_t(x)}{\pi^\varepsilon(x)} \tag{6.1.40}$$

$$= -\varepsilon \pi^\varepsilon(x) \nabla \frac{\rho_t(x)}{\pi^\varepsilon(x)} \tag{6.1.41}$$

$$= -\varepsilon \rho_t(x) \nabla \log \frac{\rho_t(x)}{\pi^\varepsilon(x)}. \tag{6.1.42}$$

Let us now differentiate $D(\rho_t\|\pi^\varepsilon)$ with respect to time:

$$\frac{\mathrm{d}}{\mathrm{d}t} D(\rho_t\|\pi^\varepsilon) = \frac{\mathrm{d}}{\mathrm{d}t} \int_{\mathbb{R}^n} \rho_t(x) \log \frac{\rho_t(x)}{\pi^\varepsilon(x)} \,\mathrm{d}x \tag{6.1.43}$$

$$= \int_{\mathbb{R}^n} \frac{\partial}{\partial t}\rho_t(x) \cdot \log \frac{\rho_t(x)}{\pi^\varepsilon(x)} \,\mathrm{d}x + \int_{\mathbb{R}^n} \rho_t \frac{\partial}{\partial t} \Big(\log \frac{\rho_t(x)}{\pi^\varepsilon(x)} \Big) \,\mathrm{d}x. \tag{6.1.44}$$

The second integral is identically zero:

$$\int_{\mathbb{R}^n} \rho_t \frac{\partial}{\partial t}\Big(\log \frac{\rho_t(x)}{\pi^\varepsilon(x)}\Big)\,\mathrm{d}x = \int_{\mathbb{R}^n} \rho_t \frac{\partial}{\partial t}\log \rho_t(x)\,\mathrm{d}x \tag{6.1.45}$$

$$= \int_{\mathbb{R}^n} \rho_t \frac{\partial}{\partial t}\rho_t(x)\,\mathrm{d}x \tag{6.1.46}$$

$$= \frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^n} \rho_t(x)\,\mathrm{d}x \tag{6.1.47}$$

$$= 0. \tag{6.1.48}$$

Thus, using (6.1.38) and integrating by parts, we can write

$$\frac{\mathrm{d}}{\mathrm{d}t}D(\rho_t\|\pi^\varepsilon) = \int_{\mathbb{R}^n} \frac{\partial}{\partial t}\rho_t(x)\cdot \log\frac{\rho_t(x)}{\pi^\varepsilon(x)}\,\mathrm{d}x \tag{6.1.49}$$

$$= -\int_{\mathbb{R}^n} \Big(\nabla\cdot j_t(x)\Big)\log\frac{\rho_t(x)}{\pi^\varepsilon(x)}\,\mathrm{d}x \tag{6.1.50}$$

$$= \int_{\mathbb{R}^n} j_t(x)^T\nabla \log\frac{\rho_t(x)}{\pi^\varepsilon(x)}\,\mathrm{d}x \tag{6.1.51}$$

$$= -\varepsilon\int_{\mathbb{R}^n} \rho_t(x)\Big|\nabla\log\frac{\rho_t(x)}{\pi^\varepsilon(x)}\Big|^2\,\mathrm{d}x. \tag{6.1.52}$$

Finally, using the definition of the relative Fisher information, we obtain the following important formula:

$$\frac{\mathrm{d}}{\mathrm{d}t}D(\rho_t\|\pi^\varepsilon) = -\varepsilon I(\rho_t\|\pi^\varepsilon). \tag{6.1.53}$$

This shows that the relative entropy $D(\rho_t\|\pi^\varepsilon)$ between the density ρ_t of X_t and the equilibrium density π^ε decreases with t; however, just as in the case of gradient descent, further conditions are needed on π^ε to ensure that it converges to 0. One such sufficient condition is as follows: There exists a constant $c>0$, such that

$$D(\rho\|\pi^\varepsilon) \le \frac{c\varepsilon}{2} I(\rho\|\pi^\varepsilon) \tag{6.1.54}$$

for all densities ρ. When (6.1.54) holds, we say that π^ε satisfies a *logarithmic Sobolev inequality* with constant c. For example, if $f(x)=\frac{1}{2}|x|^2$, so that $\pi^\varepsilon(x)\propto e^{-\frac{1}{2\varepsilon}|x|^2}$, (6.1.54) holds with $c=1$. So, if π^ε satisfies the log-Sobolev inequality with constant c, then

$$\frac{\mathrm{d}}{\mathrm{d}t}D(\rho_t\|\pi^\varepsilon) \le -\frac{2}{c}D(\rho_t\|\pi^\varepsilon), \tag{6.1.55}$$

and in this case the relative entropy between ρ_t and π^ε decays exponentially:

$$D(\rho_t\|\pi^\varepsilon) \le e^{-2t/c}D(\rho_0\|\pi^\varepsilon). \tag{6.1.56}$$

6.1.3 Simulated annealing in continuous time

As mentioned earlier, the expected value

$$\bar{f}^\varepsilon := \int_{\mathbb{R}^n} f(x)\pi^\varepsilon(x)\,\mathrm{d}x = \frac{1}{Z^\varepsilon}\int_{\mathbb{R}^n} f(x)e^{-\frac{1}{\varepsilon}f(x)}\,\mathrm{d}x \tag{6.1.57}$$

converges to the global minimum of f as the temperature parameter $\varepsilon \to 0$. Moreover, for all sufficiently large t (e.g., $t \gg c$, where c is the log-Sobolev constant of π^ε), the probability density ρ_t of X_t will be sufficiently close to π^ε. This suggests that we could eventually approach the minimum value of f by running a Langevin dynamics with time-dependent temperature $\varepsilon(t)$:

$$\mathrm{d}X_t = -\nabla f(X_t)\,\mathrm{d}t + \sqrt{2\varepsilon(t)}\,\mathrm{d}W_t. \tag{6.1.58}$$

The idea is that, on the one hand, the temperature $\varepsilon(t)$ would vary sufficiently slowly with time so that $\mathbf{E}[f(X_t)] \approx \bar{f}^{\varepsilon(t)}$ for t large enough, but would also monotonically converge to 0 as $t \to \infty$, so that

$$\bar{f}^{\varepsilon(t)} \to \bar{f}^0 \equiv \min f, \qquad \text{as } t \to \infty. \tag{6.1.59}$$

This is the basis of so-called *simulated annealing*, first proposed in the context of discrete-time optimization by Kirkpatrick, Gelatt, and Vecchi in 1983. The term "annealing" refers to a process in metallurgy where an alloy is first heated and then slowly cooled in order to increase its hardness. The rate of decrease of $\varepsilon(t)$ is called the *cooling schedule* or *protocol*.

We can reason about this heuristically as follows. Let the cooling protocol be given with $\varepsilon(0) > 0$, and let π_t denote the density $\pi^{\varepsilon(t)}$. Then we have the following for the time derivative of the relative entropy $D(\rho_t\|\pi_t)$, where ρ_t is the density of X_t in (6.1.59):

$$\frac{\mathrm{d}}{\mathrm{d}t}D(\rho_t\|\pi_t) = \int_{\mathbb{R}^n} \frac{\partial}{\partial t}\rho_t(x) \cdot \log\frac{\rho_t(x)}{\pi_t(x)}\,\mathrm{d}x + \int_{\mathbb{R}^n} \rho_t(x)\frac{\partial}{\partial t}\log\frac{\rho_t(x)}{\pi_t(x)}\,\mathrm{d}x \tag{6.1.60}$$

$$= -\varepsilon(t) I(\rho_t\|\pi_t) - \int_{\mathbb{R}^n} \rho_t(x)\frac{\partial}{\partial t}\log\pi_t(x)\,\mathrm{d}x. \tag{6.1.61}$$

To further analyze the second integral, it is convenient to work with the inverse temperature $\beta(t) := \varepsilon^{-1}(t)$. Then we can write

$$\frac{\partial}{\partial t}\log\pi_t(x) = \dot{\beta}(t)\frac{\partial}{\partial\beta}\log\pi^{1/\beta}(x)\Big|_{\beta=\beta(t)}, \tag{6.1.62}$$

and further

$$\frac{\partial}{\partial\beta}\log\pi^{1/\beta}(x) = -f(x) - \frac{\mathrm{d}}{\mathrm{d}\beta}\log Z^{1/\beta} \tag{6.1.63}$$

$$= \bar{f}^{1/\beta} - f(x) \tag{6.1.64}$$

Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}t}D(\rho_t\|\pi_t) = -\frac{1}{\beta(t)}I(\rho_t\|\pi_t) + \dot{\beta}(t)\int_{\mathbb{R}^n}\big(f(x) - \bar{f}^{1/\beta}\big)\rho_t(x)\,\mathrm{d}x. \tag{6.1.65}$$

Now suppose that each π^ε satisfies a log-Sobolev inequality, and let $c(t)$ denote the log-Sobolev constant π_t. In that case, we can further estimate

$$\frac{\mathrm{d}}{\mathrm{d}t}D(\rho_t\|\pi_t) \le -\frac{2}{c(t)}D(\rho_t\|\pi_t) + \dot{\beta}(t)\mathbf{E}[f(X_t) - \bar{f}^{1/\beta(t)}]. \tag{6.1.66}$$

Suppose that there exists a constant $C > 0$ that depends on f and n, such that

$$\mathbf{E}[f(X_t) - \bar{f}^{1/\beta(t)}] \leq \frac{C}{\beta(t)} D(\rho_t \| \pi_t) \tag{6.1.67}$$

(this may be shown in many cases provided f is sufficiently well-behaved). Then we want to choose $\dot{\beta}(t)$ such that

$$\frac{1}{c(t)} \lesssim \frac{\dot{\beta}(t)}{\beta(t)}, \tag{6.1.68}$$

at least for all t large enough. In many cases, the log-Sobolev constant of π^ε will scale like $e^{-c/\varepsilon} = e^{c\beta}$ for some constant $c > 0$. We can then take $\beta(t) \sim \log(1+t)$, which suggests taking $\beta(t) \sim \log t$ for all t sufficiently large.

6.2 Constrained minimization and stochastic neural nets

So far, we have focused on unconstrained optimization. When constraints are present, one has to make appropriate modifications to the gradient descent method to make sure that the points along its trajectory satisfy the constraints. An interesting approach was suggested by J. Hopfield. Suppose we wish to minimize a function $f : \mathbb{R}^n \to \mathbb{R}$ subject to the constraint $x \in [0,1]^n$. Consider the following continuous-time dynamics:

$$\dot{x}(t) = -\nabla f(y(t)) \tag{6.2.1a}$$
$$y(t) = \boldsymbol{\sigma}(x(t)), \tag{6.2.1b}$$

where $\boldsymbol{\sigma} : \mathbb{R}^n \to [0,1]^n$ has the form

$$\boldsymbol{\sigma}(x^1, \ldots, x^n) := \big(\sigma(x^1), \ldots, \sigma(x^n)\big)^T \tag{6.2.2}$$

for some strictly increasing, continuous function $\sigma : \mathbb{R} \to [0,1]$. For example, we could take

$$\sigma(v) = \frac{1}{2}\Big(1 + \tanh\Big(\frac{v}{a}\Big)\Big), \tag{6.2.3}$$

where $a > 0$ is a tunable parameter. We can think of (6.2.1) as an autonomous dynamical system with state $x(t) \in \mathbb{R}^n$ and output $y(t) \in \mathbb{R}^n$. The ODE (6.2.1a) can be expressed purely in terms of $x(t)$ as

$$\dot{x}(t) = -\nabla f(\boldsymbol{\sigma}(x(t))). \tag{6.2.4}$$

The effect of $\boldsymbol{\sigma}$ in (6.2.1) is to "squish" each coordinate of $x(t)$ to the interval $[0,1]$, thus ensuring that each point along the output trajectory $y(t)$ satisfies the constraints. Moreover, it is not hard to see that $f(y(t))$ is a decreasing function of t:

$$\frac{\mathrm{d}}{\mathrm{d}t} f(y(t)) = \nabla f(y(t))^T \dot{x}(t) \tag{6.2.5}$$
$$= -|\nabla f(y(t))|^2 \tag{6.2.6}$$
$$\leq 0. \tag{6.2.7}$$

However, just as in the unconstrained case, the trajectory $y(t)$ can get stuck near a local minimum. Hence, it makes sense to look for a suitable stochastic relaxation, just like we had done in the unconstrained case with the Langevin dynamics. One approach, which we will describe next, was suggested by E. Wong.

Let us modify the deterministic dynamics (6.2.1) as follows:

$$\mathrm{d}X_t = -\nabla f(Y_t)\,\mathrm{d}t + \alpha(X_t)\,\mathrm{d}W_t \tag{6.2.8a}$$

$$Y_t = \boldsymbol{\sigma}(X_t) \tag{6.2.8b}$$

where the matrix-valued function $\alpha : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ is to be chosen so that the output process $(Y_t)_{t\geq 0}$ is a diffusion process with equilibrium density given by

$$\pi^\varepsilon(y) = \frac{1}{Z^\varepsilon} \exp\Big(-\frac{1}{\varepsilon} f(y)\Big) \mathbf{1}_{\{y\in[0,1]^n\}}, \tag{6.2.9}$$

where the normalization constant Z^ε now takes the form

$$Z^\varepsilon = \int_{[0,1]^n} e^{-\frac{1}{\varepsilon}f(y)}\,\mathrm{d}y. \tag{6.2.10}$$

(Of course, the process (Y_t) is confined to the cube $[0,1]^n$ due to the effect of $\boldsymbol{\sigma}$.) Let us assume that σ is twice differentiable; since it is strictly increasing, its derivative σ' is positive everywhere. We will now show that the following simple choice gives us what we want:

$$\alpha^{ij}(x) = \begin{cases} \sqrt{\dfrac{2\varepsilon}{\sigma'(x^i)}}, & i = j \\ 0, & \text{else} \end{cases} \tag{6.2.11}$$

Suppose then that the (as yet undetermined) matrix $\alpha(x)$ is diagonal, and let $\alpha^i(x)$ denote its entry in position (i,i). If (Y_t) is a diffusion process, then it will satisfy an SDE

$$\mathrm{d}Y_t = m(Y_t)\,\mathrm{d}t + \beta(Y_t)\,\mathrm{d}W_t, \tag{6.2.12}$$

for some drift $m : \mathbb{R}^n \to \mathbb{R}^n$ and diffusion matrix $\beta : \mathbb{R}^n \to \mathbb{R}^{n\times n}$. Let us only consider the case when $\beta(y)$ is a diagonal matrix (later we will see that we will have this by construction). We first determine the conditions that m and β must satisfy to give us the equilibrium density we want. We will neglect the multiplier $\mathbf{1}_{\{y\in[0,1]^n\}}$ in (6.2.9) since $Y_t \in [0,1]^n$ by construction. Since $\beta(y)$ is diagonal, the Fokker–Planck equation for (6.2.12) takes the form

$$\frac{\partial}{\partial t}\rho_t(y) = -\sum_{i=1}^n \frac{\partial}{\partial y_i}\left(m^i(y)\rho_t(y) - \frac{1}{2}\frac{\partial}{\partial y_i}\big(B^i(y)\rho_t(y)\big)\right), \tag{6.2.13}$$

where $B(y) := \beta(y)^2$. let us denote the ith coordinate of $\nabla f(x)$ by $\partial_i f(x)$. If we want $\pi^\varepsilon(y) \propto e^{-(1/\varepsilon)f(y)}$ to be an equilibrum density, m and B must be such that

$$m^i(y)e^{-\frac{1}{\varepsilon}f(y)} - \frac{1}{2}\frac{\partial}{\partial y_i}\big(B^i(y)e^{-\frac{1}{\varepsilon}f(y)}\big) = 0, \qquad i = 1,\dots,n \tag{6.2.14}$$

which simplifies to

$$m^i(y) = \frac{1}{2}\left(-\frac{1}{\varepsilon} B^i(y)\partial_i f(y) + \frac{1}{2}\frac{\partial}{\partial y_i} B^i(y) \right), \qquad i = 1, \dots, n \tag{6.2.15}$$

Now, Itô's rule gives

$$\mathrm{d}Y_t^i = \sigma'(X_t^i)\,\mathrm{d}X_t^i + \frac{1}{2}A^i(X_t)\sigma''(X_t)\,\mathrm{d}t \tag{6.2.16}$$

$$= \left(-\partial_i f(Y_t)\sigma'(X_t^i) + \frac{1}{2}\sigma''(X_t^i)A^i(X_t) \right)\mathrm{d}t + \sigma'(X_t^i)\alpha^i(X_t)\,\mathrm{d}W_t^i, \tag{6.2.17}$$

where $A(x) := \alpha(x)^2$. Then, for each i we should have

$$\beta^i(\boldsymbol{\sigma}(x)) = \sigma'(x^i)\alpha^i(x), \tag{6.2.18}$$

so that $B^i(\boldsymbol{\sigma}(x)) = (\sigma'(x^i))^2 A^i(x)$ and therefore

$$m^i(y) = -\partial_i f(y)\sigma'(x^i) + \frac{1}{2}\frac{\sigma''(x^i)}{(\sigma'(x^i))^2} B^i(y). \tag{6.2.19}$$

Let us now define the function $\check{\sigma}(v) := \sigma'(\sigma^{-1}(v))$—recall that σ is strictly increasing and thus invertible. Then it is readily verified that

$$(\log \check{\sigma})'(v) = \frac{\sigma''(\sigma^{-1}(v))}{(\sigma'(\sigma^{-1}(v)))^2}, \tag{6.2.20}$$

and, since $x^i = \sigma^{-1}(y^i)$, we can write

$$m^i(y) = -\partial_i f(y)\check{\sigma}(y^i) + \frac{1}{2}(\log \check{\sigma})'(y^i)B^i(y). \tag{6.2.21}$$

Comparing (6.2.15) and (6.2.21) gives

$$\frac{1}{2}B^i(y) = \varepsilon\check{\sigma}(y^i), \tag{6.2.22}$$

which, together with $\sigma'(x^i) = \check{\sigma}(y^i)$, leads to

$$A^i(x) = \frac{2\varepsilon\check{\sigma}(y^i)}{(\sigma'(x^i))^2} = \frac{2\varepsilon}{\sigma'(x^i)}, \tag{6.2.23}$$

as claimed. For example, with the hyperbolic tangent choice of σ, as in (6.2.3), we have

$$\sigma'(v) = \frac{1}{2a}\left(1 - \tanh^2\left(\frac{v}{a}\right)\right) = \frac{1}{2a\cosh^2\left(\frac{v}{a}\right)}, \tag{6.2.24}$$

so the stochastic dynamics of (X_t, Y_t) becomes

$$\mathrm{d}X_t^i = -\nabla f(Y_t)\,\mathrm{d}t + 2\sqrt{a\varepsilon}\cosh\left(\frac{X_t^i}{a}\right)\mathrm{d}W_t^i, \tag{6.2.25}$$

$$Y_t^i = \frac{1}{2}\left(1 + \tanh\left(\frac{X_t^i}{a}\right)\right), \qquad i = 1, \dots, n. \tag{6.2.26}$$

6.3 Free energy minimization and optimal control

In Section 6.1, the Gibbs densities (6.1.7), for $\varepsilon > 0$ sufficiently close to zero, were motivated through a stochastic relaxation of the global minimization problem for a given $f : \mathbb{R}^n \to \mathbb{R}$. However, there is also a sense in which such densities are themselves solutions of an optimization problem in the space of probability distributions. While the origin of this idea is in statistical physics, where it is known as the *Gibbs variational principle*, it has wide applicability beyond physics.

To see how the density π^ε could arise as a solution of an optimization problem, let us define the following functional on the space of all (sufficiently well-behaved) probability densities ρ on $\mathbb{R}^n$:

$$\mathsf{G}^\varepsilon(\rho) := \int_{\mathbb{R}^n} f(x)\rho(x)\,\mathrm{d}x + \varepsilon \int_{\mathbb{R}^n} \rho(x)\log\rho(x)\,\mathrm{d}x. \tag{6.3.1}$$

The first term is the expectation of f w.r.t. ρ. If we maintain the analogy with statistical physics, then f plays the role of an energy function, so the first term is the expected energy, which we will denote by $\mathsf{E}(\rho)$. The second term can be written as $-\varepsilon\mathsf{S}(\rho)$, where

$$\mathsf{S}(\rho) := -\int_{\mathbb{R}^n} \rho(x)\log\rho(x)\,\mathrm{d}x \tag{6.3.2}$$

is the *entropy* of ρ. The quantity $\mathsf{G}^\varepsilon(\rho) = \mathsf{E}(\rho) - \varepsilon\mathsf{S}(\rho)$ is then known as the *Gibbs free energy* at temperature ε. In statistical physics, it has the intepretation of the energy available for conversion to useful work, since $\mathsf{E}(\rho)$ is the total average energy, while $\varepsilon\mathsf{S}(\rho)$ is the heat dissipated into the environment. Now, using the definition (6.1.7) of π^ε and the formula for the relative entropy $D(\rho\|\pi)$, we can write

$$\mathsf{G}^\varepsilon(\rho) = \varepsilon \int_{\mathbb{R}^n} \rho(x)\log\frac{\rho(x)}{e^{-(1/\varepsilon)f(x)}}\,\mathrm{d}x \tag{6.3.3}$$

$$= \varepsilon D(\rho\|\pi^\varepsilon) - \varepsilon\log Z^\varepsilon. \tag{6.3.4}$$

Since $D(\rho\|\pi^\varepsilon) \geq 0$, $\mathsf{G}^\varepsilon(\rho)$ is never smaller than $-\varepsilon\log Z^\varepsilon$; moreover, this minimum value is attained if and only if $\rho = \pi^\varepsilon$.

Now we will discuss a simple modification of this construction that brings many benefits. Instead of considering the negative entropy $-\mathsf{S}(\rho)$, let us fix a 'reference' density ρ_0 and define the functional

$$\mathsf{F}^\varepsilon(\rho) := \mathsf{E}(\rho) + \varepsilon D(\rho\|\rho_0). \tag{6.3.5}$$

This is also a free energy functional, and in fact we may view it as a Gibbs free energy if in (6.3.1) we replace f with $f - \varepsilon\log\rho_0$. A calculation similar to the one for G^ε shows that the minimum value of F^ε is equal to

$$-\varepsilon\log\int_{\mathbb{R}^n} \rho_0(x)e^{-(1/\varepsilon)f(x)} =: -\varepsilon\log\tilde{Z}^\varepsilon, \tag{6.3.6}$$

and is uniquely minimized by

$$\tilde{\pi}^\varepsilon(x) = \frac{e^{-(1/\varepsilon)f(x)}\rho_0(x)}{\int_{\mathbb{R}^n} e^{-(1/\varepsilon)f(x)}\rho_0(x)\,\mathrm{d}x}. \tag{6.3.7}$$

The advantage of using relative entropy instead of negative entropy, apart from the fact that the former is always nonnegative, is that Z^ε will be infinite if f is positive and bounded, whereas $\tilde{Z}^\varepsilon$ will be finite since the integration is performed w.r.t. the reference pdf ρ_0 rather than the Lebesgue measure on $\mathbb{R}^n$. Thus, f must grow sufficiently rapidly as $|x| \to \infty$ in order to π^ε to be well-defined, whereas $\tilde{\pi}^\varepsilon$ will exist under much milder requirements on f.

Many problems are free energy minimization in disguise. As an example, consider the topic of Bayesian inference: Let (X, Y) be a random couple, where we observe some 'evidence' Y correlated with some hidden 'state' X. If X has the prior density ρ_0 and if Y conditioned on $X = x$ has density $q(y|x)$, then, upon observing $Y = y$, we update the prior $\rho_0(x)$ to the posterior density $\rho(x|y)$, given by

$$\rho(x|y) = \frac{q(y|x)\rho_0(x)}{\int_{\mathbb{R}^n} q(y|x)\rho_0(x)\,\mathrm{d}x}. \tag{6.3.8}$$

If we define the y-dependent energy $f(x; y) := -\log q(y|x)$, then the posterior $\rho(x|y)$ has the Gibbs form

$$\rho(x|y) \propto e^{-f(x;y)}\rho_0(x), \tag{6.3.9}$$

and therefore minimizes the free energy

$$\mathsf{F}(\rho; y) := -\int_{\mathbb{R}^n} \rho(x) \log q(y|x)\,\mathrm{d}x + D(\rho\|\rho_0). \tag{6.3.10}$$

Now, even though we have formulated everything in terms of densities on $\mathbb{R}^n$, the Gibbs variational principle can be formulated more broadly as a minimization problem over probability measures on an abstract measurable space $(\Omega, \mathcal{F})$, where $\mathcal{F}$ is a given σ-algebra. To that end, we first expand the definition of the relative entropy to any pair of probability measures μ, ν on Ω:

$$D(\mu\|\nu) := \begin{cases} \int_\Omega \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\,\mathrm{d}\nu, & \mu \ll \nu \\ +\infty, & \text{otherwise} \end{cases} \tag{6.3.11}$$

where the notation $\mu \ll \nu$ indicates that μ is absolutely continuous w.r.t. ν, and $\frac{\mathrm{d}\mu}{\mathrm{d}\nu}$ denotes the Radon–Nikodym derivative of μ w.r.t. ν. It is not hard to see that this abstract definition reduces to our earlier one if μ and ν are probability measures on $\mathbb{R}^n$ that have densities ρ and π. At any rate, let us now fix a measurable function $f : \Omega \to \mathbb{R}$ and a reference probability measure ν on Ω and and define the free energy

$$\mathsf{F}^\varepsilon(\mu) := \mathbf{E}_\mu[f] + \varepsilon D(\mu\|\nu). \tag{6.3.12}$$

Exactly the same argument as before can be used to show that the minimum value of F^ε is equal to $-\varepsilon \log \mathbf{E}_\nu[e^{-(1/\varepsilon)f}]$, and is uniquely attained by $\bar{\mu}$ with

$$\frac{\mathrm{d}\bar{\mu}}{\mathrm{d}\nu} = \frac{e^{-(1/\varepsilon)f}}{\mathbf{E}_\nu[e^{-(1/\varepsilon)f}]}. \tag{6.3.13}$$

(A minimal regularity condition is needed here to ensure that the expected value $\mathbf{E}_\mu[e^{-(1/\varepsilon)f}]$ exists and is finite.) We will now consider the case of a particular Ω, namely the path space $C([0, T]; \mathbb{R}^n)$.

6.3.1 Free energy minimization in path space

Consider an n-dimensional diffusion process $(X_t)_{0\le t\le T}$ given by

$$X_t = x + \int_0^t f(X_s, s)\,\mathrm{d}s + W_t, \qquad 0 \le t \le T \tag{6.3.14}$$

and let $\boldsymbol{P}^x$ denote the probability law of its trajectory $\boldsymbol{X} = (X_t)_{0\le t\le T}$ viewed as a random element of $\Omega = C([0,T];\mathbb{R}^n)$. Let us define the following energy (or cost) function on Ω:

$$H(\boldsymbol{X}) := \int_0^T q_0(X_t, t)\,\mathrm{d}t + r(X_T), \tag{6.3.15}$$

where $q_0 : \mathbb{R}^n \times [0,T] \to \mathbb{R}$ and $r : \mathbb{R}^n \times \mathbb{R}$ are some given functions such that

$$Z(x) := \mathbf{E}_{\boldsymbol{P}^x}\left[e^{-H(\boldsymbol{X})}\right] < \infty. \tag{6.3.16}$$

(this will be the case, e.g., if both q_0 and r are bounded). We would like to minimize the free energy

$$\mathsf{F}(\boldsymbol{Q}^x) := \mathbf{E}_{\boldsymbol{Q}^x}[H] + D(\boldsymbol{Q}^x \| \boldsymbol{P}^x) \tag{6.3.17}$$

over all probability measures $\boldsymbol{Q}^x$ on Ω, for which $X_0 = x$ almost surely. From our discussion above, it follows that the minimizing probability measure $\bar{\boldsymbol{Q}}^x$ is given by

$$\frac{\mathrm{d}\bar{\boldsymbol{Q}}^x}{\mathrm{d}\boldsymbol{P}^x} = \frac{\exp\big(-H\big)}{Z(x)}, \tag{6.3.18}$$

which amounts to reweighting the paths $\boldsymbol{X} \sim \boldsymbol{P}^x$ using importance weights proportional to $e^{-H(\boldsymbol{X})}$— that is, for any bounded measurable function $\varphi : \Omega \to \mathbb{R}$,

$$\mathbf{E}_{\bar{\boldsymbol{Q}}^x}[\varphi(\boldsymbol{X})] = \frac{1}{Z(x)}\mathbf{E}_{\boldsymbol{P}^x}\left[\varphi(\boldsymbol{X})e^{-H(\boldsymbol{X})}\right]. \tag{6.3.19}$$

While this could, in principle, be done, we will now show that we can instead generate samples from $\bar{\boldsymbol{Q}}^x$ directly by adding a drift to (6.3.14), which is what we would expect on the basis of Gisranov's theorem. Moreover, we will see that this drift can be Viewed as a solution to a certain optimal control problem.

In what follows, we will denote by $\mathbf{E}^x[\cdot]$ (respectively, $\bar{\mathbf{E}}^x[\cdot]$) expectation w.r.t. $\boldsymbol{P}^x$ (respectively, $\bar{\boldsymbol{Q}}^x$). Since $\bar{\boldsymbol{Q}}^x$ is absolutely continuous w.r.t. $\boldsymbol{P}^x$, by Girsanov's theorem there exists an adapted drift process $\bar{\boldsymbol{u}} = (\bar{u}_t)_{0\le t\le T}$, such that

$$\frac{\mathrm{d}\bar{\boldsymbol{Q}}^x}{\mathrm{d}\boldsymbol{P}^x} = \exp\left(\int_0^T \bar{u}_t^T\,\mathrm{d}W_t - \frac{1}{2}\int_0^T |\bar{u}_t|^2\,\mathrm{d}t\right), \tag{6.3.20}$$

and thus the process

$$\bar{X}_t = x + \int_0^t \big(f(\bar{X}_s, s) + \bar{u}_s\big)\,\mathrm{d}s + \bar{W}_t, \qquad 0 \le t \le T \tag{6.3.21}$$

where $\bar{W}_t$ is a standard n-dimensional Brownian motion, will have the desired law $\bar{\boldsymbol{Q}}^x$.

We will now look for the drift of the form $\bar{u}_t = -\nabla V(X_t, t)$, where $V : \mathbb{R}^n \times [0, T] \to \mathbb{R}$ is a $C^{2,1}$ function to be determined. To see why this ansatz makes sense, let us considder an arbitrary such V. Then, with X_t defined according to (6.3.14), Itô's rule giVes

$$V(X_T, T) = V(x, 0) + \int_0^T \Big(\dot{V}(X_t, t) + \mathcal{A}V(X_t, t) \Big) \, \mathrm{d}t + \int_0^T \nabla V(X_t, t)^T \, \mathrm{d}W_t, \tag{6.3.22}$$

where $\mathcal{A} = f^T \nabla + \frac{1}{2}\Delta$ is the infinitesimal generator of (6.3.14). Thus,

$$\begin{aligned} &- \int_0^T \nabla V(X_t, t)^T \, \mathrm{d}W_t - \frac{1}{2} \int_0^T |\nabla V(X_t, t)|^2 \, \mathrm{d}t \\ &= V(x, 0) - V(X_T, T) + \int_0^T \Big(\dot{V}(X_t, t) + \mathcal{A}V(X_t, t) - \frac{1}{2} |\nabla V(X_t, t)|^2 \Big) \, \mathrm{d}t, \end{aligned} \tag{6.3.23}$$

which suggests that the function V must be such that

$$\begin{aligned} &V(x, 0) - V(X_T, T) + \int_0^T \Big(\dot{V}(X_t, t) + \mathcal{A}V(X_t, t) - \frac{1}{2} |\nabla V(X_t, t)|^2 \Big) \, \mathrm{d}t \\ &\quad = -\log Z(x) - \int_0^T q_0(X_t, t) \, \mathrm{d}t - r(X_T), \end{aligned} \tag{6.3.24}$$

or, gathering all the integrals on one side,

$$\begin{aligned} &\int_0^T \Big(\dot{V}(X_t, t) + \mathcal{A}V(X_t, t) + q_0(X_t, t) - \frac{1}{2} |\nabla V(X_t, t)|^2 \Big) \, \mathrm{d}t \\ &\quad = -\big(V(x, 0) - \log Z(x) \big) + \big(V(X_T, T) - r(X_T) \big). \end{aligned} \tag{6.3.25}$$

The best-case scenario is that both sides of the above equation are identically zero. This will be the case if the function V simultaneously obeys the following:

1. It should solve the PDE

$$\frac{\partial}{\partial t} V(x, t) + \mathcal{A}V(x, t) + q_0(x, t) = \frac{1}{2} |\nabla V(x, t)|^2, \qquad (x, t) \in \mathbb{R}^n \times [0, T] \tag{6.3.26}$$

   subject to the terminal condition $V(x, T) = r(x)$.

2. $V(x, 0) = -\log Z(x)$.

Let us now show that the function

$$V(x, t) = -\log \mathbf{E} \left[\exp \left(-r(X_T) - \int_t^T q_0(X_s, s) \, \mathrm{d}s \right) \middle| X_t = x \right] \tag{6.3.27}$$

fulfills both of these requirements. First of all, it is clear that $V(x, T) = r(x)$ and $V(x, 0) = -\log Z(x)$. It remains to show that V in (6.3.27) solves (6.3.26). To that end, we will make use of the logarithmic transformation $V(x, t) = -\log h(x, t)$. It is then a matter of straightforward calculus to show that, if V solves (6.3.26), then h solves the *linear* PDE

$$\frac{\partial}{\partial t} h(x, t) + \mathcal{A}h(x, t) - q_0(x, t) h(x, t) = 0, \qquad (x, t) \in \mathbb{R}^n \times [0, T] \tag{6.3.28}$$

with the terminal condition $h(x,T) = e^{-r(x)}$. The Feynman–Kac formula then gives the following expression for $h(x,t)$:

$$h(x,t) = \mathbf{E}\left[\exp\left(-r(X_T) - \int_t^T q_0(X_s,s)\,\mathrm{d}s \right) \middle| X_t = x \right]. \tag{6.3.29}$$

The expression for (6.3.27) follows immediately.

6.3.2 The optimal control interpretation

The PDE (6.3.26) is, in fact, a Hamilton–Jacobi–Bellman equation for the following optimal control problem:

$$\text{minimize } J(x;\boldsymbol{u}) := \mathbf{E}\left[\int_0^T \Big(\frac{1}{2}|u_t|^2 + q_0(X_t,t) \Big)\,\mathrm{d}t + r(X_T) \middle| X_0 = x \right] \tag{6.3.30}$$

over all adapted controls $\boldsymbol{u} = (u_t)_{0\le t\le T}$ subject to

$$\mathrm{d}X_t = \big(f(X_t) + u_t \big)\,\mathrm{d}t + \mathrm{d}W_t, \qquad 0 \le t \le T. \tag{6.3.31}$$

To see this, use (5.3.47) to write

$$\begin{aligned} &\mathcal{A}V(x,t) + q_0(x,t) - \frac{1}{2}|\nabla V(x,t)|^2 \\ &\quad = \min_{u\in\mathbb{R}^n} \Big\{ (f(x)+u)^{\mathsf{T}} \nabla V(x,t) + \frac{1}{2}\Delta V(x,t) + q_0(x,t) + \frac{1}{2}|u|^2 \Big\} \qquad (6.3.32) \\ &\quad = \min_{u\in\mathbb{R}^n} \Big\{ \mathcal{A}^u V(x,t) + q(x,u,t) \Big\}, \qquad (6.3.33) \end{aligned}$$

where $\mathcal{A}^u = (f+u)^{\mathsf{T}}\nabla + \frac{1}{2}\Delta$ is the infinitesimal generator of the controlled diffusion (6.3.31) and where we have defined the time-dependent state-action cost

$$q(x,u,t) := q_0(x,t) + \frac{1}{2}|u|^2. \tag{6.3.34}$$

Then we can rewrite (6.3.26) as

$$\frac{\partial}{\partial t}V(x,t) + \min_{u\in\mathbb{R}^n} \Big\{ \mathcal{A}^u V(x,t) + q(x,u,t) \Big\} = 0, \qquad (x,t) \in \mathbb{R}^n \times [0,T] \tag{6.3.35}$$

with the terminal condition $V(x,T) = r(x)$, which is the HJB equation we were after. Consequently, we obtain the following variational formula:

$$\begin{aligned} &-\log \mathbf{E}^x\left[\exp\left(-r(X_T) - \int_0^T q_0(X_t,t)\,\mathrm{d}t \right) \right] \\ &\quad = \min_{\boldsymbol{u}} \mathbf{E}\left[\int_0^T \Big(\frac{1}{2}|u_t|^2 + q_0(X_t,t) \Big)\,\mathrm{d}t + r(X_T) \middle| X_0 = x \right], \end{aligned} \tag{6.3.36}$$

where the expectation on the left-hand side is taken w.r.t. the reference process (X_t) in (6.3.14), while the right-hand side involves minimization, over all admissible controls, of the expected cost $J(\boldsymbol{u})$ w.r.t. the controlled process (6.3.31). Also note that the expression on the left-hand side has the $-\log \mathbf{E}[\exp(\cdot)]$ form, while on the right-hand side there is no exponentiation inside the expectation.

Chapter 7

Sampling and generative models

We started Chapter 2 with a discussion of the stochastic realization problem, where the goal is to represent a given random object X as a deterministic function of some 'latent' or 'internal' random object W. This is the idea behind the concept of a strong solution of an SDE: Subject to appropriate regularity conditions on the drift f and on the diffusion matrix g, the process

$$X_t = x + \int_0^t f(X_s)\,\mathrm{d}s + \int_0^t g(X_s)\,\mathrm{d}W_s \tag{7.0.1}$$

can be viewed as the output of a causal system with initial condition x_0 and Brownian motion input (W_t). (A more general version of the problem is to find, for a given process $(Y_t)_{0\le t\le T}$, the functions f, g, and h, such that $Y_t = h(X_t)$.) We can also use this as a recipe for generating *sample trajectories* of diffusion processes with prescribed f and g.

In our discussion of the Schrödinger bridge problem in Chapter 5, we were interested in obtaining a sample from a given probability density p on $\mathbb{R}^n$ by controlling a diffusion process over a finite time horizon, from $t = 0$ to $t = 1$. This is also an instance of a stochastic realization problem, where we look for a system that takes a Brownian motion $(W_t)_{0\le t\le 1}$ and produces a sample X from p while requiring minimum effort: Among all processes of the form

$$X_t = \int_0^t u_s\,\mathrm{d}s + W_t, \qquad 0 \le t \le 1 \tag{7.0.2}$$

with $X_1 \sim p$, we select the one where the expected value

$$\frac{1}{2}\mathbf{E}\left[\int_0^1 |u_t|^2\,\mathrm{d}t\right] \tag{7.0.3}$$

is the smallest. A more general variant of the Schrödinger bridge entails starting from a random initial condition $X_0 \sim p_0$ and finding a drift process $(u_t)_{0\le t\le 1}$, such that

$$X_1 = X_0 + \int_0^1 u_t\,\mathrm{d}t + W_1 \tag{7.0.4}$$

has a given density p_1 and (7.0.3) is, again, minimized.

This formulation is related to the controllability problem mentioned at the beginning of Chapter 5: Given a deterministic controlled system $\dot{x} = f(x, u)$ with input $u(t) \in \mathbb{R}^n$ and state $x(t) \in \mathbb{R}^n$, find

a control $u : [0,1] \to \mathbb{R}^m$ to transfer the system from a given initial state $x(0) = x_0$ to a given final state $x(1) = x_1$ while minimizing the control cost $\frac{1}{2}\int_0^1 |u(t)|^2 \, \mathrm{d}t$. The difference is that, in the stochastic case, we are interested in optimally controlling a diffusion process to go from a given initial density p_0 to a given final density p_1.

7.1 Generative modeling

This control perspective can be brought to bear on the problem of generative modeling: Given a target density p, construct a diffusion process $(X_t)_{0 \le t \le T}$, such that X_0 has some reference density q (we can think of it as a 'prior') and X_T has the target density p. We can think of this in two ways:

1. Controlling the initial density $\rho_0 = q$ to the target density $\rho_T = p$ subject to a suitable optimality criterion. This is an optimal control problem in the space of densities.

2. Realizing $(X_t)_{0 \le t \le T}$ as a strong solution of some SDE such that X_T will have the target density p when X_0 has the prior density q, again subject to a suitable optimality criterion.

In a sense, Schrödinger's bridge already gives us a framework for this. However, the solution for random (as opposed to deterministic) initial conditions cannot be given in closed form. Moreover, instead of full knowledge of the target density p, we may have access to a large number of independent samples $X_0^1, \dots, X_0^N \sim p$. Thus, we have to settle for suboptimal solutions that are learned on the basis of these samples. In this chapter, we will take a look at one particularly effective approach to this, referred to as *diffusion models* in the machine learning communiy.

Here is the basic idea. Let us first consider the setting when the target p is known; we will return to the sample-based version later. Suppose that we have a diffusion process governed by an SDE

$$\mathrm{d}X_t = f(X_t, t) \, \mathrm{d}t + \sigma(t) \, \mathrm{d}W_t, \tag{7.1.1}$$

where the drift f and the diffusion coefficient σ are such that, if we were to initialize this process with a sample X_0 from the target p and run it for some time T, the resulting density ρ_T of X_T would be very close to the prior q. For example, we could simply take $f \equiv 0$, in which case the density of

$$X_T = X_0 + \int_0^T \sigma(t) \, \mathrm{d}W_t, \tag{7.1.2}$$

with X_0 independent of the Brownian motion (W_t), is given by the convolution of ρ_0 and a Gaussian density:

$$\rho_T(x) = \rho_0 * \gamma_{\Sigma(T)}(x) = \frac{1}{(2\pi\Sigma(T))^{n/2}} \int_{\mathbb{R}^n} \rho_0(\xi) \exp\left(-\frac{1}{2\Sigma^2(T)}|x-\xi|^2\right) \mathrm{d}\xi, \tag{7.1.3}$$

where $\Sigma^2(T) := \int_0^T \sigma^2(t) \, \mathrm{d}t$. If $\sigma(t)$ increases with t quickly enough, then we can take q to be the Gaussian density $\gamma_{\Sigma(T)}$. Another choice is $f(x,t) = -\frac{\sigma^2(t)}{2}x$, which gives the time-inhomogeneous Ornstein–Uhlenbeck process

$$\mathrm{d}X_t = -\frac{\sigma^2(t)}{2} X_t \, \mathrm{d}t + \sigma(t) \, \mathrm{d}W_t. \tag{7.1.4}$$

Then we have

$$X_T = \exp\left(-\frac{1}{2}\int_0^T \sigma^2(t)\,\mathrm{d}t\right) X_0 + \int_0^T \exp\left(-\frac{1}{2}\int_t^T \sigma^2(s)\,\mathrm{d}s\right)\sigma(t)\,\mathrm{d}W_s$$
$$= \exp\Big(-\frac{1}{2}\Sigma^2(T)\Big)X_0 + \exp\Big(-\frac{1}{2}\Sigma^2(T)\Big)\int_0^T \exp\Big(\frac{1}{2}\Sigma^2(t)\Big)\sigma(t)\,\mathrm{d}W_t, \tag{7.1.5}$$

so, if T is sufficiently large, we can take q to be the standard n-dimensional Gaussian density $\gamma(x) = \frac{1}{(2\pi)^{n/2}} e^{-|x|^2/2}$. What these two examples have in common is that both f and σ are of very simple form completely unrelated to the target density p, and the overall effect is to add enough noise to X_0 so that the contribution of p to the resulting density at time T is, for all practical purposes, negligible. Thus, we now have the means of taking a sample from p and transforming it into an approximate sample from q. It is useful to think of this as an approximate multidimensional analogue of the procedure where we take a sample X from a given univariate probability distribution and convert it into a sample from the uniform distribution on $[0,1]$ by means of the transformation $X \mapsto F_X(X)$, where F_X is the cdf of X. The next step is to develop a diffusion process analogue of the reverse step $U \mapsto F_X^{-1}(U)$ of generating a sample X using the inverse cdf F_X^{-1}. The natural candidate for this is the time-reversed version of $(X_t)_{0\le t\le T}$, i.e., the process $\bar{X}_t := X_{T-t}$. Of course, in order to arrive at p at time T we would have to start with ρ_T, not with q. Here, as before, we rely on the fact that ρ_T is very close to q, so once again we are settling for an approximation.

7.2 Time reversal of diffusions

Consider an n-dimensional diffusion process governed by an Itô SDE

$$\mathrm{d}X_t = f(X_t,t)\,\mathrm{d}t + \sigma(t)\,\mathrm{d}W_t, \qquad 0 \le t \le T \tag{7.2.1}$$

where the drift depends on both space and time, while the diffusion coefficient depends only on time. If X_0 has density ρ_0, then the evolution of the density ρ_t of X_t is described by the Fokker–Planck (or Kolmogorov's forward) equation

$$\frac{\partial}{\partial t}\rho_t(x) = -\nabla\cdot\big(f(x,t)\rho_t(x)\big) + \frac{\sigma^2(t)}{2}\Delta\rho_t(x), \qquad (x,t)\in\mathbb{R}^n\times[0,T]. \tag{7.2.2}$$

The time-reversed process $\bar{X}_t$ is defined for $t\in[0,T]$ by $\bar{X}_t := X_{T-t}$. It is not hard to show that, since X_t is a Markov process, so is $\bar{X}_t$. A more interesting result is that, under some regularity conditions on ρ_0, f, and σ, $\bar{X}_t$ will also be a *diffusion process*. We give a heuristic derivation in lieu of a rigorous treatment (see Notes at the end of the chapter for references). We will assume that f and σ are sufficiently regular so that (7.2.1) has a unique strong solution, and that the initial density ρ_0 is such that (7.2.2) has a solution which is everywhere positive.

Since $\bar{X}_t$ has density $\bar{\rho}_t := \rho_{T-t}$, we have

$$\frac{\partial}{\partial t}\bar{\rho}_t(x) = \nabla\cdot\big(f(x,T-t)\bar{\rho}_t(x)\big) - \frac{\sigma^2(T-t)}{2}\Delta\bar{\rho}_t(x). \tag{7.2.3}$$

This almost looks like a Fokker–Planck equation except for the minus sign in front of the Laplacian. However, since $\Delta = \nabla\cdot\nabla$, we can add and subtract

$$\sigma^2(T-t)\Delta\bar{\rho}_t(x) = \sigma^2(T-t)\nabla\cdot\nabla\bar{\rho}_t(x) = \sigma^2(T-t)\nabla\cdot\big(\bar{\rho}_t(x)\nabla\log\bar{\rho}_t(x)\big) \tag{7.2.4}$$

on the right-hand side to get

$$\frac{\partial}{\partial t}\bar{\rho}_t(x) = -\nabla \cdot \Big(\big(-f(x, T-t) + \sigma^2(T-t)\nabla \log \bar{\rho}_t(x) \big)\bar{\rho}_t(x) \Big) + \frac{\sigma^2(T-t)}{2}\Delta\bar{\rho}_t(x). \tag{7.2.5}$$

We can now read the drift and the diffusion coefficients of $\bar{X}_t$ directly off (7.2.5):

$$\bar{f}(x,t) := -f(x, T-t) + \sigma^2(T-t)\nabla \log \rho_{T-t}(x) \tag{7.2.6a}$$
$$\bar{\sigma}(t) := \sigma(T-t). \tag{7.2.6b}$$

This gives us the following SDE for $\bar{X}_t$:

$$\mathrm{d}\bar{X}_t = \bar{f}(\bar{X}_t, t)\,\mathrm{d}t + \bar{\sigma}(t)\,\mathrm{d}\bar{W}_t, \qquad 0 \le t \le T. \tag{7.2.7}$$

The form of $\bar{f}$ in (7.2.6a) deserves further comments. Let us first consider the deterministic case $\sigma(t) \equiv 0$. Then all the randomness comes from the initial condition $X_0 \sim \rho_0$, and the density ρ_t obeys the PDE

$$\frac{\partial}{\partial t}\rho_t(x) = -\nabla \cdot \big(f(x,t)\rho_t(x) \big), \tag{7.2.8}$$

often referred to as the *Liouville equation*, which does not have any second-order terms on the right-hand side. In this instance, $\bar{f}(x,t) = -f(x, T-t)$, which is consistent with the easily checked fact that the time reversal $\bar{x}(t) := x(T-t)$ of the deterministic dynamics

$$\frac{\mathrm{d}}{\mathrm{d}t}x(t) = f(x(t), t), \qquad 0 \le t \le T \tag{7.2.9}$$

is described by

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{x}(t) = -f(\bar{x}(t, T-t)), \qquad 0 \le t \le T. \tag{7.2.10}$$

When $\sigma(t) > 0$, though, ordinary time reversal is not enough, and we pick up a second term proportional to $\nabla \log \rho_{T-t}(x)$, which is called the *score function* (or just the score) of $\rho_{T-t}(x)$. For now, it will suffice to think of it as the additional force we have to apply in order to counteract the stochastic effects of the Brownian motion. Later on, we will endow this additional term with an operational interpretation through the lens of stochastic thermodynamics.

7.2.1 An optimal control interpretation

An alternative way to obtain the drift $\bar{f}(x,t)$ in (7.2.6a) is to consider a controlled diffusion

$$\mathrm{d}\bar{X}^{\boldsymbol{u}}_t = \big(\bar{f}^0(\bar{X}^{\boldsymbol{u}}_t, t)\,\mathrm{d}t + \bar{\sigma}(t)u_t \big)\,\mathrm{d}t + \bar{\sigma}(t)\,\mathrm{d}\bar{W}_t, \qquad 0 \le t \le T \tag{7.2.11}$$

where $\bar{f}^0(x,t) := -f(x, T-t)$, and $\boldsymbol{u} = (u_t)_{0 \le t \le T}$ is an n-dimensional control (with all the usual admissibilty caveats applied). When there is no control, i.e., $u_t \equiv 0$ for all t, we obtain what we might call the *naive* time reversal, i.e., the process

$$\bar{X}^0_t = \bar{X}^0_0 + \int_0^t \bar{f}^0(\bar{X}^0_s, s)\,\mathrm{d}s + \int_0^t \bar{\sigma}(s)\,\mathrm{d}\bar{W}_s \tag{7.2.12}$$

that results if we don't account for the effect of the Brownian motion and just go with the drift that we read off the time-reversed Liouville equation. The catch is that, even if we initialize it with $\bar{X}^0 \sim \rho_T$, we cannot expect the resulting densities $\bar{\rho}_t^0$ of $\bar{X}_t^0$ to track the desired time-reversed densities $\bar{\rho}_t = \rho_{T-t}$. This provides the rationale for introducing the control term into (7.2.11). As we show next, this problem is an instance of free energy minimization of the type considered in Section 6.3. Note, however, that Eq. (7.2.11) has a time-dependent diffusion coefficient, whereas (6.3.31) has $\sigma(t)] \equiv 1$.

As a starting point, let us recall the PDE (7.2.3). Using the chain rule, the divergence term on the right-hand side can be written as

$$\nabla \cdot \big(f(x, T-t)\bar{\rho}_t(x)\big) = \big(\nabla \cdot f(x, T-t)\big)\bar{\rho}_t(x) + f(x, T-t)^T \nabla \bar{\rho}_t(x) \tag{7.2.13}$$

$$\equiv -\big(\nabla \cdot \bar{f}^0(x,t)\big)\bar{\rho}_t(x) - \bar{f}^0(x,t)^T \nabla \bar{\rho}_t(x), \tag{7.2.14}$$

which, upon rearranging, gives

$$\frac{\partial}{\partial t}\bar{\rho}_t(x) + \bar{f}^0(x,t)^T \nabla \bar{\rho}_t(x) + \frac{\bar{\sigma}^2(t)}{2}\Delta \bar{\rho}_t(x) + \big(\nabla \cdot \bar{f}^0(x,t)\big)\bar{\rho}_t(x) = 0. \tag{7.2.15}$$

Recognizing the second-order differential operator $\mathcal{A}_t^0 := (\bar{f}^0)^T \nabla + \frac{\bar{\sigma}^2(t)}{2}\Delta$ as the infinitesimal generator of the uncontrolled process $\bar{X}_t^0$, we can express (7.2.15) in the following more suggestive form:

$$\frac{\partial}{\partial t}\bar{\rho}_t(x) + \mathcal{A}_t^0 \bar{\rho}_t(x) + \big(\nabla \cdot \bar{f}^0(x,t)\big)\bar{\rho}_t(x) = 0, \quad (x,t) \in \mathbb{R}^n \times [0,T] \tag{7.2.16}$$

with the terminal condition $\bar{\rho}_T(x) = \rho_0(x)$. Written like this, (7.2.16) has the same form as (6.3.26), and an application of the Feynman–Kac formula gives the following path-space representation of $\bar{\rho}_t(x)$:

$$\bar{\rho}_t(x) = \mathbf{E}\left[\rho_0(\bar{X}_T^0) \exp\left(\int_t^T \big(\nabla \cdot \bar{f}^0(\bar{X}_s^0, s)\big)\,\mathrm{d}s\right)\middle| \bar{X}_t^0 = x\right] \tag{7.2.17}$$

We can now proceed in the same manner as in Section 6.3.2 to obtain an optimal control formulation of the time-reversal problem.

Specifically, if we define the running cost

$$q(x,u,t) := \frac{1}{2}|u|^2 - \nabla \cdot \bar{f}^0(x,t) \tag{7.2.18}$$

and the terminal cost

$$r(x) := -\log \rho_0(x), \tag{7.2.19}$$

then we readily obtain the representation of $-\log \bar{\rho}_t(x)$ as the value function

$$V(x,t) := \min_{\boldsymbol{u}} \mathbf{E}\left[\int_t^T q(\bar{X}_s^{\boldsymbol{u}}, u_s, s)\,\mathrm{d}s + \rho(\bar{X}_T^{\boldsymbol{u}})\middle| \bar{X}_t^{\boldsymbol{u}} = x\right]. \tag{7.2.20}$$

Moreover, the value function satisfies the HJB equation

$$\frac{\partial}{\partial t}V(x,t) + \min_{u \in \mathbb{R}^n}\Big\{\mathcal{A}_t^u V(x,t) + q(x,u,t)\Big\} = 0, \qquad (x,t) \in \mathbb{R}^n \times [0,T] \tag{7.2.21}$$

with the terminal conditon $V(x,t) = r(x) = -\log \rho_0(x)$, where

$$\mathcal{A}_t^u := \left(\bar{f}^0\right)^T \nabla + \bar{\sigma}(t) u^T \nabla + \frac{\bar{\sigma}^2(t)}{2}\Delta \tag{7.2.22}$$

for each $u \in \mathbb{R}^n$. Carrying out the minimization explicitly gives

$$\frac{\partial}{\partial t}V(x,t) + \mathcal{A}_t^0 V(x,t) - \nabla \cdot \bar{f}^0(x,t) = \frac{\bar{\sigma}^2(t)}{2}|\nabla V(x,t)|^2, \qquad (x,t) \in \mathbb{R}^n \times [0,T] \tag{7.2.23}$$

with $V(x,T) = -\log \rho_0(x)$, and the optimal control is given by

$$u_t = -\bar{\sigma}(t)\nabla V(x,t) = \bar{\sigma}(t)\nabla \log \rho_{T-t}(x), \tag{7.2.24}$$

i.e., $\bar{\sigma}(t)u_t$ matches exactly the extra score-dependent term in (7.2.6a).

7.2.2 Error analysis

The main benefit of the control-theoretic interpretation is that we can now quanfity the errors incurred due to (a) initializing the time-reversed process with a density different from $\bar{\rho}_0 = \rho_T$ and (b) using a suboptimal feedback control $k(x,t)$ instead of $\bar{\sigma}(t)\nabla \log \rho_{T-t}(x)$.

We start with the latter. It will be convenient to write $k(x,t)$ as $\bar{\sigma}(t)\hat{k}(x,t)$. Let $\bar{\boldsymbol{P}}^x$ denote the probability law of the process

$$\bar{X}_t^x = x + \int_0^t \bar{f}(\bar{X}_s^x, s)\,\mathrm{d}s + \int_0^t \bar{\sigma}(s)\,\mathrm{d}\bar{W}_s \tag{7.2.25}$$

$$= x + \int_0^t \left(\bar{f}^0(\bar{X}_s^x, s) + \bar{\sigma}(s)\nabla \rho_{T-s}(\bar{X}_s^x)\right)\mathrm{d}s + \int_0^t \bar{\sigma}(s)\,\mathrm{d}\bar{W}_s \qquad 0 \le t \le T \tag{7.2.26}$$

i.e., of $\bar{X}_t$ conditioned on $\bar{X}_0 = x$, and let $\hat{\boldsymbol{P}}^x$ denote the probability law of the process

$$\hat{X}_t^x = x + \int_0^t \left(\bar{f}^0(\hat{X}_s^x, s) + \bar{\sigma}(s)\hat{k}(\hat{X}_s^x, s)\right)\mathrm{d}s + \int_0^t \bar{\sigma}(s)\,\mathrm{d}\bar{W}_s, \qquad 0 \le t \le T. \tag{7.2.27}$$

Since these two processes differ only in their drift terms, Girsanov's theorem gives

$$\frac{\mathrm{d}\hat{\boldsymbol{P}}^x}{\mathrm{d}\bar{\boldsymbol{P}}^x} = \exp\left(\int_0^T \left(\hat{k}(\bar{X}_t^x, t) - \nabla \log \rho_{T-t}(\bar{X}_t^x)\right)^T \mathrm{d}\bar{W}_t - \frac{1}{2}\int_0^T \left|\hat{k}(\bar{X}_t^x, t) - \nabla \log \rho_{T-t}(\bar{X}_t^x)\right|^2 \mathrm{d}t\right). \tag{7.2.28}$$

Thus, we have the following for the relative entropy between $\bar{\boldsymbol{P}}^x$ and $\hat{\boldsymbol{P}}^x$:

$$D(\bar{\boldsymbol{P}}^x \| \hat{\boldsymbol{P}}^x) = \int \mathrm{d}\bar{\boldsymbol{P}}^x \log \frac{\mathrm{d}\bar{\boldsymbol{P}}^x}{\mathrm{d}\hat{\boldsymbol{P}}^x} \tag{7.2.29}$$

$$= -\int \mathrm{d}\bar{\boldsymbol{P}}^x \log \frac{\mathrm{d}\hat{\boldsymbol{P}}^x}{\mathrm{d}\bar{\boldsymbol{P}}^x} \tag{7.2.30}$$

$$= \mathbf{E}\left[\int_0^T \frac{1}{2}\left|\hat{k}(\bar{X}_t^x, t) - \nabla \log \rho_{T-t}(\bar{X}_t^x)\right|^2 \mathrm{d}t\right]. \tag{7.2.31}$$

Now let $\bar{\boldsymbol{P}}$ denote the probability law of $(\bar{X}_t)_{0\le t\le T}$ with $\bar{X}_0 \sim \rho_T$ and let $\hat{\boldsymbol{P}}$ denote the probability law of $(\hat{X}_t)_{0\le t\le T}$ with $\hat{X}_0 \sim q$, i.e.,

$$\bar{\boldsymbol{P}} = \int_{\mathbb{R}^n} \rho_T(x)\bar{\boldsymbol{P}}^x \,\mathrm{d}x, \qquad \hat{\boldsymbol{P}} = \int_{\mathbb{R}^n} q(x)\hat{\boldsymbol{P}}^x \,\mathrm{d}x. \tag{7.2.32}$$

Then we can use the chain rule for the relative entropy to write

$$\begin{aligned} D(\bar{\boldsymbol{P}}\|\hat{\boldsymbol{P}}) &= D(\rho_T\|q) + \int_{\mathbb{R}^n} \rho_T(x) D(\bar{\boldsymbol{P}}^x\|\hat{\boldsymbol{P}}^x) \\ &= D(\rho_T\|q) + \int_{\mathbb{R}^n} \rho_T(x)\mathbf{E}\left[\int_0^T \frac{1}{2}\left|\hat{k}(\bar{X}_t^x,t) - \nabla\log\rho_{T-t}(\bar{X}_t^x)\right|^2 \mathrm{d}t\right]\mathrm{d}x \\ &= D(\rho_T\|q) + \frac{1}{2}\int_0^T \mathbf{E}\left[\left|\hat{k}(\bar{X}_t,t) - \nabla\log\rho_{T-t}(\bar{X}_t)\right|^2\right]\mathrm{d}t. \end{aligned} \tag{7.2.33}$$

The first term on the right-hand side quantifies the error due to misspecification of the initial density, ρ_T vs. q, while the second term quantifies the error due to using a suboptimal feedback control. Moreover, if we denote by $\hat{\rho}_T$ the density of $\hat{X}_T$, then by the data processing inequality we have

$$D(\rho_0\|\hat{\rho}_T) \le D(\rho_T\|q) + \frac{1}{2}\int_0^T \mathbf{E}\left[\left|\hat{k}(\bar{X}_t,t) - \nabla\log\rho_{T-t}(\bar{X}_t)\right|^2\right]\mathrm{d}t. \tag{7.2.34}$$

As an example, let us consider the time-reversed version of the Ornstein–Uhlenbeck process in (7.1.4) when initialized with $X_0 \sim \rho_0$, i.e.,

$$\mathrm{d}\bar{X}_t = \bar{\sigma}^2(t)\Big(\frac{1}{2}\bar{X}_t + \nabla\log\rho_{T-t}(\bar{X}_t)\Big)\,\mathrm{d}t + \bar{\sigma}(t)\,\mathrm{d}\bar{W}_t, \qquad 0\le t\le T \tag{7.2.35}$$

where ρ_t is the density of

$$X_t = e^{-\frac{1}{2}\Sigma^2(t)}X_0 + \int_0^t e^{-\frac{1}{2}\left(\Sigma^2(t)-\Sigma^2(s)\right)}\sigma(s)\,\mathrm{d}W_s \tag{7.2.36}$$

with $\Sigma^2(t) := \int_0^t \sigma^2(s)\,\mathrm{d}s$. In this instance, we have

$$f(x,t) = -\frac{\sigma^2(t)}{2}x, \qquad \bar{f}^0(x,t) = \frac{\bar{\sigma}^2(t)}{2}x \tag{7.2.37}$$

A simple calculation shows that $q = \gamma$, the standard n-dimensional Gaussian density, is the equilibrium density of (7.1.4):

$$f(x,t) - \frac{\sigma^2(t)}{2}\nabla\log q(x) = -\frac{\sigma^2(t)}{2}x - \frac{\sigma^2(t)}{2}\nabla\log\gamma(x) = 0. \tag{7.2.38}$$

Thus, exactly the same calculations as in Section 6.1.2 yield

$$\frac{\mathrm{d}}{\mathrm{d}t}D(\rho_t\|q) = \frac{\mathrm{d}}{\mathrm{d}t}D(\rho_t\|\gamma) \tag{7.2.39}$$

$$= -\frac{\sigma^2(t)}{2}I(\rho_t\|\gamma), \tag{7.2.40}$$

where $I(\rho_t\|\gamma)$ is the relative Fisher information between ρ_t and γ. Since γ satisfies the log-Sobolev inequality

$$D(\rho\|\gamma) \le \frac{1}{2} I(\rho\|\gamma), \tag{7.2.41}$$

we obtain

$$D(\rho_t\|\gamma) \le \exp\left(-\int_0^t \sigma^2(s)\,\mathrm{d}s\right) D(\rho_0\|\gamma) \tag{7.2.42}$$

$$= \exp\big(-\Sigma^2(t)\big) D(\rho_0\|\gamma). \tag{7.2.43}$$

If, for example, $\sigma(t) = \sqrt{2}$, then the initialization error term in (7.2.33), with $q = \gamma$, will scale like e^{-2T} and can therefore be made negligible for sufficiently large T. On the other hand, increasing T will also make the second term larger, so there is a trade-off between these two terms.

7.3 Sample-based construction and score matching

So far, our discussion was confined to the ideal situation when the target density p was given. However, in most applications of interest this is not the case; instead, only a large number of independent samples from p is available. This presents a difficulty since the implementation of the time-reversed process $\bar{X}_t$, even when initialized with the target-independent prior q, requires knowledge of the score functions $\nabla \log \rho_t(x)$. The solution is to learn the score from the data via the procedure known as *score matching*.

The following basic idea underlies score matching: Let ρ be a probability density on $\mathbb{R}^n$, and let $s(x)$ denote its score, i.e., $s(x) = \nabla \log \rho(x)$. For any C^1 vector-valued function $\hat{s} : \mathbb{R}^n \to \mathbb{R}^n$, let us define the score estimation error

$$L(\hat{s}) := \mathbf{E}_\rho[|s(X) - \hat{s}(X)|^2] \tag{7.3.1}$$

$$= \int_{\mathbb{R}^n} \rho(x)|s(x) - \hat{s}(x)|^2. \tag{7.3.2}$$

Expanding the squared norm, we get

$$L(\hat{s}) = \int_{\mathbb{R}^n} \rho(x)\big(|s(x)|^2 - 2s(x)^T\hat{s}(x) + |\hat{s}(x)|^2\big)\,\mathrm{d}x \tag{7.3.3}$$

$$= \int_{\mathbb{R}^n} \rho(x)|s(x)|^2\rho(x)\,\mathrm{d}x - 2\int_{\mathbb{R}^n} \rho(x)s(x)^T\hat{s}(x)\,\mathrm{d}x + \int_{\mathbb{R}^n} \rho(x)|\hat{s}(x)|^2)\,\mathrm{d}x. \tag{7.3.4}$$

The first term on the right-hand side does not depend on $\hat{s}$, so we focus on the other two. The second term is of particular interest, since it is not given by an expectation, with respect to ρ, of a function that involves only $\hat{s}$. However, we can use the fact that $s(x) = \frac{1}{\rho(x)}\nabla\rho(x)$ and integration by parts to get

$$\int_{\mathbb{R}^n} \rho(x)s(x)^T\hat{s}(x)\,\mathrm{d}x = \int_{\mathbb{R}^n} \nabla\rho(x)^T\hat{s}(x)\,\mathrm{d}x \tag{7.3.5}$$

$$= -\int_{\mathbb{R}^n} \big(\nabla\cdot\hat{s}(x)\big)\rho(x)\,\mathrm{d}x, \tag{7.3.6}$$

assuming both ρ and $\hat{s}$ are decaying sufficiently rapidly at infinity. We can thus express $L(\hat{s})$ as a sum

$$L(\hat{s}) = \tilde{L}(\hat{s}) + C, \tag{7.3.7}$$

where C is a constant that depends only on ρ and where

$$\tilde{L}(\hat{s}) := \mathbf{E}_\rho\Big[|\hat{s}(X)|^2 + 2\nabla \cdot \hat{s}(X)\Big] \tag{7.3.8}$$

can be approximated using independent samples $X_1, \dots, X_N \sim \rho$ as

$$\tilde{L}(\hat{s}) \approx \tilde{L}_N(\hat{s}) := \frac{1}{N}\sum_{i=1}^{N}\Big[|\hat{s}(X_i)|^2 + 2\nabla \cdot \hat{s}(X_i)\Big]. \tag{7.3.9}$$

We now apply these ideas to the problem of estimating the scores $\nabla \log \rho_t$. Let $\mathcal{S}$ be a family of candidate score models, e.g., neural nets $\hat{s}(\cdot, \cdot; \theta) : \mathbb{R}^n \times [0, T] \to \mathbb{R}^n$ parametrized by a vector of weights θ. Given N independent samples $X_0^1, \dots, X_0^N$ from the target density p, we generate N independent diffusion trajectories

$$X_t^i := X_0^i + \int_0^t f(X_s^i, s)\,\mathrm{d}s + \int_)^t \sigma(s)\,\mathrm{d}W_s^i, \qquad 0 \le t \le T,\ i = 1, \dots, N \tag{7.3.10}$$

and then obtain the score network weights $\hat{\theta}$ by minimizing the empirical loss

$$L_N(\theta) := \int_0^T \frac{1}{N}\sum_{i=1}^{N}\Big[|\hat{s}(X_t^i, t; \theta)|^2 + 2\nabla \cdot \hat{s}(X_t^i, t; \theta)\Big]\,\mathrm{d}t \tag{7.3.11}$$

over θ. To obtain an approximate sample $\hat{X}$ from q, we then proceed as follows: Generate a sample $\hat{X}_0$ from the prior q and then take

$$\hat{X} = \hat{X}_0 + \int_0^T \big(-f(\hat{X}_t, T-t) + \bar{\sigma}^2(T-t)\hat{s}(\hat{X}_t, T-t)\big)\,\mathrm{d}t + \int_0^T \sigma(T-t)\,\mathrm{d}\bar{W}_t, \tag{7.3.12}$$

where $\hat{s}(x, t) \equiv \hat{s}(x, t; \hat{\theta})$ is the estimated score.

7.4 Stochastic thermodynamics

The optimal control interpretation of time reversal of diffusion processes suggests that it is useful to think of the extra $\sigma^2(T-t)\nabla \log \rho_{T-t}(x)$ term in (7.2.6a) as the additional force we need to apply in order to counteract the effect of the Brownian motion. From this perspective, we are controlling the state of diffusion process, so all the theory of Chapter 5 applies. However, we have also mentioned a complementary perspective, according to which the subject of control is actually a *probability density* that evolves according to the Fokker–Planck equation. This viewpoint can be phrased as a generalization of classical thermodynamics to stochastic systems.

To fix ideas, let us consider a time-inhomogeneous n-dimensional Langevin diffusion process of the form

$$\mathrm{d}X_t = -\nabla U(X_t, t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t, \qquad 0 \le t \le T \tag{7.4.1}$$

where $U : \mathbb{R}^n \times [0,T] \to \mathbb{R}$ is a smooth (e.g., $C^{2,1}$) time-varying potential. Let X_0 be the random initial condition with density ρ_0, and let ρ_t denote, as before, the density of ρ_t. The Fokker–Planck equation

$$\frac{\partial}{\partial t}\rho_t(x) = \nabla \cdot \big(\nabla U(x,t)\rho_t(x)\big) + \Delta \rho_t(x) \tag{7.4.2}$$

governs the evolution of ρ_t viewed as a *state* of a dynamical system defined over densities. We can define two functions of state:

$$\mathsf{E}_t(\rho) := \int_{\mathbb{R}^n} \rho(x) U(x,t)\,\mathrm{d}x, \tag{7.4.3}$$

the (average) energy at time t, and

$$\mathsf{S}(\rho) := -\int_{\mathbb{R}^n} \rho(x)\log\rho(x)\,\mathrm{d}x, \tag{7.4.4}$$

the entropy. Their difference, $\mathsf{G}_t(\rho) := \mathsf{E}_t(\rho) - \mathsf{S}(\rho)$, is the *free energy*. We are interested in the free energy difference $\mathsf{G}_T(\rho_T) - \mathsf{G}_0(\rho_0)$.

Let $\ell(x,t) := \log \rho_t(x)$. Then Itô's rule gives

$$U(X_T,T) = U(X_0,0) + \int_0^T \dot{U}(X_t,t)\,\mathrm{d}t + \int_0^T \mathcal{A}_t U(X_t,t)\,\mathrm{d}t + M_T^U \tag{7.4.5}$$

and

$$\ell(X_T,T) = \ell(X_0,0) + \int_0^T \dot{\ell}(X_t,t)\,\mathrm{d}t + \int_0^T \mathcal{A}_t \ell(X_t,t)\,\mathrm{d}t + M_T^\ell, \tag{7.4.6}$$

where $\dot{U}(x,t) := \frac{\partial}{\partial t}U(x,t)$, $\dot{\ell}(x,t) := \frac{\partial}{\partial t}\ell(x,t)$, $\mathcal{A}_t := -\nabla U(\cdot,t)^T\nabla + \Delta$ is the (time-dependent) infinitesimal generator of (7.4.1), and M_T^U, M_T^ℓ are zero-mean martingales (they can be given explicitly as Itô integrals, but we will not need that). We further have

$$\mathcal{A}_t U(x,t) = -|\nabla U(x,t)|^2 + \Delta U(x,t), \tag{7.4.7}$$

$$\mathcal{A}_t \ell(x,t) = -\nabla U(x,t)^T\nabla \ell(x,t) + \Delta \ell(x,t). \tag{7.4.8}$$

Here, $\nabla \ell(x,t) \equiv \nabla \log \rho_t(x)$ is the score at time t. Moreover, using (7.4.2), we have

$$\dot{\ell}(x,t) = \frac{1}{\rho_t(x)}\frac{\partial}{\partial t}\rho_t(x) \tag{7.4.9}$$

$$= \frac{1}{\rho_t(x)}\Big(\nabla U(x,t)^T\nabla\rho_t(x) + \big(\Delta U(x,t)\rho_t(x)\big)\Big) + \frac{1}{\rho_t(x)}\Delta\rho_t(x) \tag{7.4.10}$$

$$= \nabla U(x,t)^T\nabla\ell(x,t) + \Delta U(x,t) + \frac{1}{\rho_t(x)}\Delta\rho_t(x). \tag{7.4.11}$$

Since

$$\Delta\ell(x,t) = \nabla\cdot\Big(\frac{1}{\rho_t(x)}\nabla\rho_t(x)\Big) \tag{7.4.12}$$

$$= -|\nabla\ell(x,t)|^2 + \frac{1}{\rho_t(x)}\Delta\rho_t(x), \tag{7.4.13}$$

we can write

$$\dot{\ell}(x,t) = \nabla U(x,t)^T \nabla \ell(x,t) + \Delta U(x,t) + \Delta \ell(x,t) + |\nabla \ell(x,t)|^2. \tag{7.4.14}$$

Substituting these into (7.4.5) and (7.4.6) and simplifying, we get

$$\begin{aligned} &U(X_T,T) - U(X_0,0) \\ &= \int_0^T \dot{U}(X_t,t)\,\mathrm{d}t + \int_0^T \Big(-|\nabla U(X_t,t)|^2 + \Delta U(X_t,t) \Big)\,\mathrm{d}t + M_T^U \end{aligned} \tag{7.4.15}$$

and

$$\begin{aligned} &\ell(X_T,T) - \ell(X_0,0) \\ &= \int_0^T \Big(\nabla U(X_t,t)^T \nabla \ell(X_t,t) + \Delta U(X_t,t) + \Delta \ell(X_t,t) + |\nabla \ell(X_t,t)|^2 \Big)\,\mathrm{d}t \\ &\quad + \int_0^T \Big(-\nabla U(X_t,t)^T \nabla \ell(X_t,t) + \Delta \ell(X_t,t) \Big)\,\mathrm{d}t + M_T^\ell \qquad (7.4.16) \\ &= \int_0^T \Big(\Delta U(X_t,t) + 2\Delta \ell(X_t,t) + |\nabla \ell(X_t,t)|^2 \Big)\,\mathrm{d}t + M_T^\ell. \qquad (7.4.17) \end{aligned}$$

Adding these together and taking expectations gives

$$\begin{aligned} &\mathsf{G}_T(\rho_T) - \mathsf{G}_0(\rho_0) \\ &= \int_0^T \mathbf{E}[\dot{U}(X_t,t)]\,\mathrm{d}t + \int_0^T \mathbf{E}[2\Delta U(X_t,t) + 2\Delta \ell(X_t,t) - |\nabla U(X_t,t)|^2 + |\nabla \ell(X_t,t)|^2]\,\mathrm{d}t. \end{aligned} \tag{7.4.18}$$

Let us now examine the expectation in the second integral. Using integration by parts, we have

$$\begin{aligned} \mathbf{E}[\Delta U(X_t,t)] &= \int_{\mathbb{R}^n} \rho_t(x) \Delta U(x,t)\,\mathrm{d}x && (7.4.19)\\ &= -\int_{\mathbb{R}^n} \nabla \rho_t(x)^T \nabla U(x,t)\,\mathrm{d}x && (7.4.20)\\ &= -\int_{\mathbb{R}^n} \rho_t(x) \nabla \ell(x,t)^T \nabla U(x,t)\,\mathrm{d}x && (7.4.21)\\ &= -\mathbf{E}[\nabla \ell(X_t,t)^T U \ell(X_t,t)], && (7.4.22)\\ \mathbf{E}[\Delta \ell(X_t,t)] &= \int_{\mathbb{R}^n} \rho_t(x) \Delta \ell(x,t)\,\mathrm{d}x && (7.4.23)\\ &= -\int_{\mathbb{R}^n} \nabla \rho_t(x)^T \nabla \ell(x,t)\,\mathrm{d}x && (7.4.24)\\ &= -\int_{\mathbb{R}^n} \rho_t(x) |\nabla \ell(x,t)|^2\,\mathrm{d}x && (7.4.25)\\ &= -\mathbf{E}[|\nabla \ell(X_t)|^2]. && (7.4.26) \end{aligned}$$

Thus,

$$\begin{aligned} &\mathbf{E}[2\Delta U(X_t,t) + 2\Delta \ell(X_t,t) - |\nabla U(X_t,t)|^2 + |\nabla \ell(X_t,t)|^2] \\ &\quad = -\mathbf{E}[|\nabla U(X_t,t)|^2 + 2\nabla U(X_t,t)^T \nabla \ell(X_t) + |\nabla \ell(X_t,t)|^2]. && (7.4.27)\\ &\quad = -\mathbf{E}[|\nabla U(X_t,t) + \nabla \ell(X_t,t)|^2]. && (7.4.28) \end{aligned}$$

Putting everything together and rearranging, we obtain the following relation:

$$\mathsf{G}_T(\rho_T) - \mathsf{G}_0(\rho_0) = \mathsf{W}(\rho_0 \to \rho_T) - \mathsf{D}(\rho_0 \to \rho_T), \tag{7.4.29}$$

where we have defined the expected *work*

$$\mathsf{W}(\rho_0 \to \rho_T) := \int_0^T \int_{\mathbb{R}^n} \rho_t(x) \dot{U}(x,t) \, \mathrm{d}x \, \mathrm{d}t \tag{7.4.30}$$

and expected *dissipation*

$$\mathsf{D}(\rho_0 \to \rho_T) := \int_0^T \int_{\mathbb{R}^n} \rho_t(x) |\nabla U(x,t) + \nabla \log \rho_t(x)|^2 \, \mathrm{d}x \, \mathrm{d}t, \tag{7.4.31}$$

with the notation $\rho_0 \to \rho_T$ indicating that, unlike the free energy difference which is a function of initial and final states ρ_0 and ρ_T, both W and D depend on the entire state *trajectory* $(\rho_t)_{0 \le t \le T}$.

The relation (7.4.29) is a key formula in stochastic thermodynamics. Its immediate consequence is the inequality

$$\mathsf{W}(\rho_0 \to \rho_T) \ge \mathsf{G}_T(\rho_T) - \mathsf{G}_0(\rho_0), \tag{7.4.32}$$

which can be thought of as a stochastic version of the second law of thermodynamics, which says that the expected work in effecting the transfer from one density-valued state to another can never be smaller than the difference in free energies between the final state and the initial state. Moreover, equality in (7.4.32) is attained if and only if the expected dissipation is identically zero, i.e., if and only if $\nabla \log \rho_t(x) = -\nabla U(x,t)$ for almost all x and t.

We can now apply the relation between work, free energy, and dissipation both to (7.4.1) (with $X_0 \sim \rho_0$) and to its time reversal

$$\mathrm{d}\bar{X}_t = \big(\nabla U(\bar{X}_t, T-t) + 2\nabla \log \rho_{T-t}(\bar{X}_t)\big) \, \mathrm{d}t + \sqrt{2} \, \mathrm{d}\bar{W}_t, \qquad 0 \le t \le T \tag{7.4.33}$$

which we can also express in the time-varying Langevin form as

$$\mathrm{d}\bar{X}_t = -\nabla \bar{U}(\bar{X}_t, t) \, \mathrm{d}t + \sqrt{2} \, \mathrm{d}\bar{W}_t \tag{7.4.34}$$

with $\bar{U}(x,t) := -U(x, T-t) - 2 \log \rho_{T-t}(x)$. Then, with $\bar{\rho}_t \equiv \rho_{T-t}$ denoting the density of $\bar{X}_t$, we see that

$$\mathsf{G}_T(\bar{\rho}_T) - \mathsf{G}_0(\bar{\rho}_0) = \mathsf{G}_0(\rho_0) - \mathsf{G}_T(\rho_T) \tag{7.4.35}$$

and

$$\begin{aligned}
\mathsf{D}(\bar{\rho}_0 \to \bar{\rho}_T) &= \int_0^T \int_{\mathbb{R}^n} \bar{\rho}_t(x) |\nabla \bar{U}(x,t) + \nabla \log \bar{\rho}_t(x)|^2 \, \mathrm{d}x \, \mathrm{d}t && (7.4.36)\\
&= \int_0^T \int_{\mathbb{R}^n} \rho_t(x) |\nabla U(x,t) + \nabla \log \rho_t(x)|^2 \, \mathrm{d}x \, \mathrm{d}t && (7.4.37)\\
&= \mathsf{D}(\rho_0 \to \rho_T), && (7.4.38)
\end{aligned}$$

which gives the following relation between the work done in transferring ρ_0 to ρ_T along (7.4.1) and the work done in transfering ρ_T to ρ_0 along (7.4.33):

$$\mathsf{W}(\rho_0 \to \rho_T) + \mathsf{W}(\rho_T \to \rho_0) = 2\mathsf{D}(\rho_0 \to \rho_T). \tag{7.4.39}$$

Appendix A

Probability Facts

Lemma 5 (Borel–Cantelli)**.** *Let $A_0, A_1, \ldots$ be a sequence of events.*

1. *If $\sum_{n\geq 0} \mathbf{P}[A_n] < \infty$, then*

$$\mathbf{P}\left[\limsup_{n\to\infty} A_n\right] = 0. \tag{A.0.1}$$

2. *If $\sum_{n\geq 0} \mathbf{P}[A_n] = \infty$ and the events (A_n) are independent, then*

$$\mathbf{P}\left[\limsup_{n\to\infty} A_n\right] = 1. \tag{A.0.2}$$

A.0.1 Convergence concepts

Appendix B

Analysis Facts

Lemma 6. *Let $f_0, f_1, \ldots$ be a sequence of real-valued continuous functions on the interval $[0,1]$ that converges uniformly to some function $f : [0,1] \to \mathbb{R}$, i.e.,*

$$\lim_{n\to\infty} \sup_{t\in[[0,1]} |f_n(t) - f(t)| = 0. \tag{B.0.1}$$

Then f is itself a continuous function.

Lemma 7. *Let $f_0, f_1, \ldots$ be a sequence of continuous functions on $[0,1]$, such that*

$$\sum_{n=1}^{\infty} \sup_{t\in[0,1]} |f_n(t) - f_{n-1}(t)| < \infty. \tag{B.0.2}$$

Then f_n converge uniformly to a continuous function $f : [0,1] \to \mathbb{R}$.

Bibliography

[Ben71] Vaclav E. Beneš. Existence of optimal stochastic control laws. *SIAM Journal on Control*, 9(3):446–472, August 1971.

[Buc04] James Antonio Bucklew. *Introduction to Rare Event Simulation*. Springer, 2004.

[Cla66] J. M. C. Clark. *The Representation of Non-Linear Stochastic Systems with Applications to Filtering*. PhD thesis, Electrical Engineering Department, Imperial College, London, 1966.

[Cla73] J. M. C. Clark. An introduction to stochastic differential equations on manifolds. In D. Q. Mayne and R. W. Brockett, editors, *Geometric Methods in System Theory*, pages 131–149. Reidel, Dordrecht, Holland, 1973.

[CM44] R. H. Cameron and W. T. Martin. Transformations of Wiener integrals under translations. *Annals of Mathematics*, 45:386–396, 1944.

[Doo53] J. L. Doob. *Stochastic Processes*. Wiley, 1953.

[FH65] Richard P. Feynman and Arthur R. Hibbs. *Quantum Mechanics and Path Integrals*. McGraw-Hill, 1965.

[FKK72] M. Fujisaki, G. Kallianpur, and H. Kunita. Stochastic differential equations for the nonlinear filtering problem. *Osaka Journal of Mathematics*, 9:19–40, 1972.

[Gir60] Igor V. Girsanov. On transforming a certainn class of stochastic processes by absolutely continuous substitution of measures. *Theory of Probability and Its Applications*, 5:285–301, 1960.

[Kac49] Mark Kac. On distributions of certain wiener functionals. *Transactions of American Mathematicall Society*, 65:1–13, 1949.

[Kac85] Mark Kac. *Enigmas of Chance: An Autobiography*. University of California Press, Berkeley, CA, 1985.

[Kam81] N. G. van Kampen. Itô vs. Stratonovich. *Journal of Statistical Physics*, 24(1):175–187, 1981.

[KS98] Ioannis Karatzas and Steven E. Shreve. *Brownian Motion and Stochastic Calculus*. Springer, 2nd edition, 1998.

[Mor69] Richard E. Mortensen. Mathematical problems of modeling stochastic nonlinear dynamical systems. *Journal of Statistical Physics*, 1(2):271–296, 1969.

[Nel67] E. Nelson. *Dynamical Theories of Brownian Motion.* Princeton University Press, 1967.

[Pic91] G. Picci. Stochastic realization theory. In A. C. Antoulas, editor, *Mathematical System Theory: The Influence of R. E. Kalman*, pages 213–229. Springer, 1991.

[Roe94] Gert Roepstorff. *Path Integral Approach to Quantum Physics.* Springer, 1994.

[Sim05] Barry Simon. *Functional Integration and Quantum Physics.* American Mathematical Society, 2nd edition, 2005.

[Ste01] J. Michael Steele. *Stochastic Calculus and Financial Applications.* Springer, 2001.

[SW74] Jan H. van Schuppen and Eugene Wong. Transformation of local martingales under a change of measure. *Annals of Probability*, 2(5):879–888, 1974.

[Wil89] Jan C. Willems. Models for dynamics. In U. Kirchgraber and H. O. Walther, editors, *Dynamics Reported*, volume 2, pages 171–269. Wiley, 1989.

[Won73] Eugene Wong. Recent progress in stochastic processes – A survey. *IEEE Transactions on Information Theory*, IT-19(3):262–275, May 1973.

[WZ65a] Eugene Wong and Moshe Zakai. On the convergence of ordinary integrals to stochastic integrals. *Annals of Mathematical Statistics*, 46:1560–1564, 1965.

[WZ65b] Eugene Wong and Moshe Zakai. On the relation between ordinary and stochastic differential equations. *International Journal of Engineering Science*, 3:213–229, 1965.