Written material for the course held in Autumn 2014
Version 1.1 (December 4, 2014)

Lecture Notes on **Applied Stochastic Differential Equations**

Version as of December 4, 2014

Simo Särkkä and Arno Solin

Preface

The purpose of these notes is to provide an introduction to stochastic differential equations (SDEs) from an applied point of view. The main application described is Bayesian inference in SDE models, including Bayesian filtering, smoothing, and parameter estimation. However, we have also included some SDE examples arising in physics and electrical engineering. Because the aim is in applications, much more emphasis is put into solution methods than to analysis of the theoretical properties of the equations. From pedagogical point of view the purpose of these notes is to provide an intuitive understanding in what SDEs are all about, and if the reader wishes to learn the formal theory later, he/she can read, for example, the brilliant books of Øksendal (2003) and Karatzas and Shreve (1991).

The pedagogical aim is also to overcome one slight disadvantage in many SDE books (*e.g.*, the above-mentioned ones), which is that they lean heavily on measure theory, rigorous probability theory, and to the theory martingales. There is nothing wrong in these theories—they are very powerful theories and everyone should indeed master them. However, when these theories are explicitly used in explaining SDEs, a lot of technical details need to be taken care of. When studying SDEs for the first time this tends to blur the basic ideas and intuition behind the theory. In these notes, with no shame, we trade rigour to readability when treating SDEs completely without measure theory.

In these notes, the aim is also to present an overview of numerical approximation methods for SDEs. Along with the Itô–Taylor series based simulation methods and stochastic Runge–Kutta methods the overview covers Gaussian approximation based methods which have and are still used a lot in the context of optimal filtering. Application of these methods to Bayesian inference in SDEs is presented as well.

The final chapter of the notes is currently a less organized collection of important topics including brief descriptions of martingale properties of SDEs, Girsanov theorem, Feynman–Kac formulas, series expansions, as well as Fourier domain methods for SDEs.

This is now the third version of these notes and for the first time Arno is on board as well. We are going to improve these notes in the next versions. If/when you find mistakes, please feel free to report them to us.

Best regards,
Simo and Arno

Contents

Chapter 1

Some background on ordinary differential equations

1.1 What is an ordinary differential equation?

An ordinary differential equation (ODE) is an equation, where the unknown quantity is a function, and the equation involves derivatives of the unknown function. For example, the second order differential equation for a forced spring (or, *e.g.*, a resonator circuit in telecommunications) can be generally expressed as

$$\frac{\mathrm{d}^2 x(t)}{\mathrm{d}t^2} + \gamma \frac{\mathrm{d}x(t)}{\mathrm{d}t} + \nu^2 x(t) = w(t), \tag{1.1}$$

where ν and γ are constants which determine the resonant angular velocity and damping of the spring. The force $w(t)$ is some given function which may or may not depend on time. In this equation the position variable x is called the *dependent variable* and time t is the *independent variable*. The equation is of *second order*, because it contains the second derivative and it is *linear*, because $x(t)$ appears linearly in the equation. The equation is inhomogeneous, because it contains the forcing term $w(t)$. This inhomogeneous term will become essential in later chapters, because replacing it with a random process leads to a stochastic differential equation.

Here the *solution* to the differential equation is defined as a *particular solution*, where it satisfies the equation and does not contain any arbitrary constants. A *general solution* on the other hand contains every particular solution of the equation parametrized by some free constants. To actually solve the differential equation it is necessary tie down the general solution by the initial conditions of the differential equation. In the above case it means that we need to know the spring position $x(t_0)$ and velocity $\mathrm{d}x(t_0)/\mathrm{d}t$ at some fixed initial time t_0. Given these initial values, there is a unique solution to the equation (provided that $w(t)$ is continuous). Instead of initial conditions, we could also fix some other (boundary) conditions of the differential equation to get a unique solution, but here we shall only consider differential equations with given initial conditions.

Note that it is common not to write the dependencies of x and w on t explicitly, and write the equation as

$$\frac{\mathrm{d}^2 x}{\mathrm{d}t^2} + \gamma \frac{\mathrm{d}x}{\mathrm{d}t} + \nu^2 x = w. \tag{1.2}$$

Although this sometimes is misleading, this "ink saving" notation is very commonly used and we shall also employ it here whenever there is no risk of confusion. Furthermore, because in these notes we only consider ordinary differential equations, we often drop the word "ordinary" and just talk about differential equations.

Time derivatives are also sometimes denoted with dots over the variable such as $\dot{x} = \mathrm{d}x/\mathrm{d}t$, $\ddot{x} = \mathrm{d}^2x/\mathrm{d}t^2$, and so on. In this Newtonian notation the above differential equation would be written as

$$\ddot{x} + \gamma \dot{x} + \nu^2 x = w. \tag{1.3}$$

Differential equations of an arbitrary order n can (almost) always be converted into vector differential equations of order one. For example, in the spring model above, if we define a *state variable* as $\mathbf{x}(t) = (x_1, x_2) = (x(t), \mathrm{d}x(t)/\mathrm{d}t)$, we can rewrite the above differential equation as first order vector differential equation as follows:

$$\underbrace{\begin{pmatrix} \mathrm{d}x_1(t)/\,\mathrm{d}t \\ \mathrm{d}x_2(t)/\,\mathrm{d}t \end{pmatrix}}_{\mathrm{d}\mathbf{x}(t)/\mathrm{d}t} = \underbrace{\begin{pmatrix} 0 & 1 \\ -\nu^2 & -\gamma \end{pmatrix} \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix}}_{\mathbf{f}(\mathbf{x}(t))} + \underbrace{\begin{pmatrix} 0 \\ 1 \end{pmatrix}}_{\mathbf{L}} w(t). \tag{1.4}$$

The above equation can be seen to be a special case of models of the form

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}(t), t) + \mathbf{L}(\mathbf{x}(t), t)\, \mathbf{w}(t), \tag{1.5}$$

where the vector valued function $\mathbf{x}(t) \in \mathbb{R}^n$ is generally called the state of the system, and $\mathbf{w}(t) \in \mathbb{R}^s$ is some (vector valued) forcing function, driving function or input to the system. Note that we can absorb the second term on the right to the first term to yield

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}(t), t), \tag{1.6}$$

and in that sense Equation (1.5) is slightly redundant. However, the form (1.5) turns out to be useful in the context of stochastic differential equations and thus it is useful to consider it explicitly.

The first order vector differential equation representation of an nth differential equation is often called state-space form of the differential equation. Because nth order differential equations can always be converted into equivalent vector valued first order differential equations, it is convenient to just consider such first order equations instead of considering nth order equations explicitly. Thus in these notes we develop the theory and solution methods only for first order vector differential equations and assume that nth order equations are always first converted into equations of this class.

The spring model in Equation (1.4) is also a special case of *linear differential equations* of the form

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \mathbf{F}(t)\,\mathbf{x}(t) + \mathbf{L}(t)\,\mathbf{w}(t), \tag{1.7}$$

which is a very useful class of differential equations often arising in applications. The usefulness of linear equations is that we can actually solve these equations unlike general non-linear differential equations. This kind of equations will be analyzed in the next section.

1.2 Solutions of linear time-invariant differential equations

Consider the following scalar linear homogeneous differential equation with a fixed initial condition at $t = 0$:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f\,x, \qquad x(0) = \text{given}, \tag{1.8}$$

where f is a constant. This equation can now be solved, for example, via separation of variables, which in this case means that we formally multiply by $\mathrm{d}t$ and divide by x to yield

$$\frac{\mathrm{d}x}{x} = f\ \mathrm{d}t. \tag{1.9}$$

If we now integrate left hand side from $x(0)$ to $x(t)$ and right hand side from 0 to t, we get

$$\ln x(t) - \ln x(0) = f\,t, \tag{1.10}$$

or

$$x(t) = \exp(f\,t)\,x(0). \tag{1.11}$$

Another way of arriving to the same solution is by integrating both sides of the original differential equation from 0 to t. Because $\int_0^t \mathrm{d}x/\mathrm{d}t\ \ \mathrm{d}t = x(t) - x(0)$, we can then express the solution $x(t)$ as

$$x(t) = x(0) + \int_0^t f\,x(\tau)\,\mathrm{d}\tau. \tag{1.12}$$

We can now substitute the right hand side of the equation for x inside the integral, which gives:

$$\begin{aligned} x(t) &= x(0) + \int_0^t f\,\left[x(0) + \int_0^\tau f\,x(\tau)\ \mathrm{d}\tau\right]\mathrm{d}\tau \\ &= x(0) + f\,x(0)\int_0^t \mathrm{d}\tau + \iint_0^t f^2\,x(\tau)\ \mathrm{d}\tau^2 \\ &= x(0) + f\,x(0)\,t + \iint_0^t f^2\,x(\tau)\ \mathrm{d}\tau^2. \end{aligned} \tag{1.13}$$

Doing the same substitution for $x(t)$ inside the last integral further yields

$$\begin{aligned} x(t) &= x(0) + f\,x(t0)\,t + \iint_0^t f^2\,[x(0) + \int_0^\tau f\,x(\tau)\,\mathrm{d}\tau]\,\mathrm{d}\tau^2 \\ &= x(0) + f\,x(0)\,t + f^2\,x(0)\iint_0^t \mathrm{d}\tau^2 + \iiint_0^t f^3\,x(\tau)\,\mathrm{d}\tau^3 \\ &= x(0) + f\,x(0)\,t + f^2\,x(0)\,\frac{t^2}{2} + \iiint_0^t f^3\,x(\tau)\,\mathrm{d}\tau^3. \end{aligned} \tag{1.14}$$

It is easy to see that repeating this procedure yields to the solution of the form

$$\begin{aligned} x(t) &= x(0) + f\,x(0)\,t + f^2\,x(0)\,\frac{t^2}{2} + f^3\,x(0)\,\frac{t^3}{6} + \ldots \\ &= \left(1 + f\,t + \frac{f^2\,t^2}{2!} + \frac{f^3\,t^3}{3!} + \ldots\right) x(0). \end{aligned} \tag{1.15}$$

The series in the parentheses can be recognized to be the Taylor series for $\exp(f\,t)$. Thus, provided that the series actually converges (it does), we again arrive at the solution

$$x(t) = \exp(f\,t)\,x(0) \tag{1.16}$$

The multidimensional generalization of the homogeneous linear differential equation (1.8) is an equation of the form

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{F}\,\mathbf{x}, \qquad \mathbf{x}(0) = \text{given}, \tag{1.17}$$

where $\mathbf{F}$ is a constant (time-independent) matrix. For this multidimensional equation we cannot use the separation of variables method, because it only works for scalar equations. However, the second series based approach indeed works and yields to a solution of the form

$$\mathbf{x}(t) = \left(\mathbf{I} + \mathbf{F}\,t + \frac{\mathbf{F}^2\,t^2}{2!} + \frac{\mathbf{F}^3\,t^3}{3!} + \ldots\right) \mathbf{x}(0). \tag{1.18}$$

The series in the parentheses now can be seen to be a matrix generalization of the exponential function. This series indeed is the definition of the matrix exponential:

$$\exp(\mathbf{F}\,t) = \mathbf{I} + \mathbf{F}\,t + \frac{\mathbf{F}^2\,t^2}{2!} + \frac{\mathbf{F}^3\,t^3}{3!} + \ldots \tag{1.19}$$

and thus the solution to Equation (1.17) can be written as

$$\mathbf{x}(t) = \exp(\mathbf{F}\,t)\,\mathbf{x}(0). \tag{1.20}$$

Note that the matrix exponential cannot be computed by computing scalar exponentials of the individual elements in matrix $\mathbf{F}\,t$, but it is a completely different

function. Sometimes the matrix exponential is written as expm($\mathbf{F}\,t$) to distinguish it from the elementwise computation definition, but here we shall use the common convention to simply write it as exp($\mathbf{F}\,t$). The matrix exponential function can be found as a built-in function in most commercial and open source mathematical software packages. In addition to this kind of numerical solution, the exponential can be evaluated analytically, for example, by directly using the Taylor series expansion, by using the Laplace or Fourier transform, or via the Cayley–Hamilton theorem (Åström and Wittenmark, 1996).

Example 1.1 (Matrix exponential)**.** *To illustrate the difference of the matrix exponential and elementwise exponential, consider the equation*

$$\frac{\mathrm{d}^2 x}{\mathrm{d}t^2} = 0, \quad x(0) = \textit{given}, \quad (\mathrm{d}x/\mathrm{d}t)(0) = \textit{given}, \tag{1.21}$$

which in state space form can be written as

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \underbrace{\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}}_{\mathbf{F}} \mathbf{x}, \qquad \mathbf{x}(0) = \textit{given}, \tag{1.22}$$

where $\mathbf{x} = (x, \mathrm{d}x/\mathrm{d}t)$*. Because* $\mathbf{F}^n = \mathbf{0}$ *for* $n > 1$*, the matrix exponential is simply*

$$\exp(\mathbf{F}\,t) = \mathbf{I} + \mathbf{F}\,t = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} \tag{1.23}$$

which is completely different from the elementwise matrix:

$$\begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} \neq \begin{pmatrix} \exp(0) & \exp(1) \\ \exp(0) & \exp(0) \end{pmatrix} = \begin{pmatrix} 1 & e \\ 1 & 1 \end{pmatrix} \tag{1.24}$$

Let's now consider the following linear differential equation with an inhomogeneous term on the right hand side:

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \mathbf{F}\,\mathbf{x}(t) + \mathbf{L}\,\mathbf{w}(t), \tag{1.25}$$

where $\mathbf{x}(t_0)$ is given and the matrices $\mathbf{F}$ and $\mathbf{L}$ are constant. For inhomogeneous equations the solution methods are quite few especially if we do not want to restrict ourselves to specific kinds of forcing functions $\mathbf{w}(t)$. However, the *integrating factor* method can be used for solving generic inhomogeneous equations.

Let's now move the term $\mathbf{F}\,\mathbf{x}$ to the left hand side and multiply with a magical term called integrating factor $\exp(-\mathbf{F}\,t)$ which results in the following:

$$\exp(-\mathbf{F}\,t)\,\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} - \exp(-\mathbf{F}\,t)\,\mathbf{F}\,\mathbf{x}(t) = \exp(-\mathbf{F}\,t)\,\mathbf{L}(t)\,\mathbf{w}(t). \tag{1.26}$$

From the definition of the matrix exponential we can derive the following property:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\exp(-\mathbf{F}\,t)\right] = -\exp(-\mathbf{F}\,t)\,\mathbf{F}. \tag{1.27}$$

The key things is now to observe that

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\exp(-\mathbf{F}\,t)\,\mathbf{x}(t)\right] = \exp(-\mathbf{F}\,t)\,\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} - \exp(-\mathbf{F}\,t)\,\mathbf{F}\,\mathbf{x}(t), \tag{1.28}$$

which is exactly the left hand side of Equation (1.26). Thus we can rewrite the equation as

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\exp(-\mathbf{F}\,t)\,\mathbf{x}(t)\right] = \exp(-\mathbf{F}\,t)\,\mathbf{L}(t)\,\mathbf{w}(t). \tag{1.29}$$

Integrating from t_0 to t then gives

$$\exp(-\mathbf{F}\,t)\,\mathbf{x}(t) - \exp(-\mathbf{F}\,t_0)\,\mathbf{x}(t_0) = \int_{t_0}^{t} \exp(-\mathbf{F}\,\tau)\,\mathbf{L}(\tau)\,\mathbf{w}(\tau)\;\mathrm{d}\tau, \tag{1.30}$$

which further simplifies to

$$\mathbf{x}(t) = \exp(\mathbf{F}\,(t-t_0))\,\mathbf{x}(t_0) + \int_{t_0}^{t} \exp(\mathbf{F}\,(t-\tau))\,\mathbf{L}(\tau)\,\mathbf{w}(\tau)\;\mathrm{d}\tau, \tag{1.31}$$

The above expression is thus the complete solution to the Equation (1.25).

1.3 Solutions of general linear differential equations

In this section we consider solutions of more general, time-varying linear differential equations. The corresponding stochastic equations are a very useful class of equations, because they can be solved in (semi-)closed form quite much analogously to the deterministic case considered in this section.

The solution presented in the previous section in terms of matrix exponential only works if the matrix F is constant. Thus for the time-varying homogeneous equation of the form

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{F}(t)\,\mathbf{x}, \quad \mathbf{x}(t_0) = \text{given}, \tag{1.32}$$

the matrix exponential solution does not work. However, we can implicitly express the solution in form

$$\mathbf{x}(t) = \mathbf{\Psi}(t, t_0)\,\mathbf{x}(t_0), \tag{1.33}$$

where $\mathbf{\Psi}(t, t_0)$ is the transition matrix which is defined via the properties

$$\begin{aligned}
\partial\mathbf{\Psi}(\tau, t)/\partial\tau &= \mathbf{F}(\tau)\,\mathbf{\Psi}(\tau, t) \\
\partial\mathbf{\Psi}(\tau, t)/\partial t &= -\mathbf{\Psi}(\tau, t)\,\mathbf{F}(t) \\
\mathbf{\Psi}(\tau, t) &= \mathbf{\Psi}(\tau, s)\,\mathbf{\Psi}(s, t) \\
\mathbf{\Psi}(t, \tau) &= \mathbf{\Psi}^{-1}(\tau, t) \\
\mathbf{\Psi}(t, t) &= \mathbf{I}.
\end{aligned} \tag{1.34}$$

Given the transition matrix we can then construct the solution to the inhomogeneous equation

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \mathbf{F}(t)\,\mathbf{x}(t) + \mathbf{L}(t)\,\mathbf{w}(t), \quad \mathbf{x}(t_0) = \text{given}, \tag{1.35}$$

analogously to the time-invariant case. This time the integrating factor is $\mathbf{\Psi}(t_0, t)$ and the resulting solution is:

$$\mathbf{x}(t) = \mathbf{\Psi}(t, t_0)\,\mathbf{x}(t_0) + \int_{t_0}^{t} \mathbf{\Psi}(t, \tau)\,\mathbf{L}(\tau)\,\mathbf{w}(\tau)\ \mathrm{d}\tau. \tag{1.36}$$

1.4 Fourier transforms

One very useful method to solve inhomogeneous linear time invariant differential equations is the Fourier transform. The *Fourier transform* of a function $g(t)$ is defined as

$$G(i\,\omega) = \mathcal{F}[g(t)] = \int_{-\infty}^{\infty} g(t)\ \exp(-i\,\omega\,t)\ \mathrm{d}t. \tag{1.37}$$

and the corresponding inverse Fourier transform is

$$g(t) = \mathcal{F}^{-1}[G(i\,\omega)] = \frac{1}{2\pi}\int_{-\infty}^{\infty} G(i\,\omega)\ \exp(-i\,\omega\,t)\ \mathrm{d}\omega. \tag{1.38}$$

The usefulness of the Fourier transform for solving differential equations arises from the property

$$\mathcal{F}[\mathrm{d}^n g(t)/\mathrm{d}t^n] = (i\,\omega)^n\,\mathcal{F}[g(t)], \tag{1.39}$$

which transforms differentiation into multiplication by $i\,\omega$, and from the convolution theorem which says that convolution gets transformed into multiplication:

$$\mathcal{F}[g(t) * h(t)] = \mathcal{F}[g(t)]\,\mathcal{F}[h(t)], \tag{1.40}$$

where the convolution is defined as

$$g(t) * h(t) = \int_{-\infty}^{\infty} g(t-\tau)\,h(\tau)\ \mathrm{d}\tau. \tag{1.41}$$

In fact, the above properties require that the initial conditions are zero. However, this is not a restriction in practice, because it is possible to tweak the inhomogeneous term such that its effect is equivalent to the given initial conditions.

To demonstrate the usefulness of Fourier transform, let's consider the spring model

$$\frac{\mathrm{d}^2 x(t)}{\mathrm{d}t^2} + \gamma\,\frac{\mathrm{d}x(t)}{\mathrm{d}t} + \nu^2\,x(t) = w(t). \tag{1.42}$$

Taking Fourier transform of the equation and using the derivative rule we get

$$(i\,\omega)^2\,X(i\,\omega) + \gamma\,(i\,\omega)\,X(i\,\omega) + \nu^2\,X(i\,\omega) = W(i\,\omega), \tag{1.43}$$

where $X(i\,\omega)$ is the Fourier transform of $x(t)$, and $W(i\,\omega)$ is the Fourier transform of $w(t)$. We can now solve for $X(i\,\omega)$ which gives

$$X(i\,\omega) = \frac{W(i\,\omega)}{(i\,\omega)^2 + \gamma\,(i\,\omega) + \nu^2} \tag{1.44}$$

The solution to the equation is then given by the inverse Fourier transform

$$x(t) = \mathcal{F}^{-1}\left[\frac{W(i\,\omega)}{(i\,\omega)^2 + \gamma\,(i\,\omega) + \nu^2}\right]. \tag{1.45}$$

However, for general $w(t)$ it is useful to note that the term on the right hand side is actually a product of the transfer function

$$H(i\,\omega) = \frac{1}{(i\,\omega)^2 + \gamma\,(i\,\omega) + \nu^2} \tag{1.46}$$

and $W(i\,\omega)$. This product can now be converted into convolution if we start by computing the impulse response function

$$\begin{aligned} h(t) &= \mathcal{F}^{-1}\left[\frac{1}{(i\,\omega)^2 + \gamma\,(i\,\omega) + \nu^2}\right] \\ &= b^{-1}\,\exp(-a\,t)\,\sin(b\,t)\,u(t), \end{aligned} \tag{1.47}$$

where $a = \gamma/2$ and $b = \sqrt{\nu^2 - \gamma^2/4}$, and $u(t)$ is the Heaviside step function, which is zero for $t < 0$ and one for $t \geq 0$. Then the full solution can then expressed as

$$x(t) = \int_{-\infty}^{\infty} h(t-\tau)\,w(\tau)\,\mathrm{d}\tau, \tag{1.48}$$

which can be interpreted such that we construct $x(t)$ by feeding the signal $w(t)$ though a linear system (filter) with impulse responses $h(t)$.

We can also use Fourier transform to solve general LTI equations

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \mathbf{F}\,\mathbf{x}(t) + \mathbf{L}\,\mathbf{w}(t). \tag{1.49}$$

Taking Fourier transform gives

$$(i\,\omega)\,\mathbf{X}(i\,\omega) = \mathbf{F}\,\mathbf{X}(i\,\omega) + \mathbf{L}\,\mathbf{W}(i\,\omega), \tag{1.50}$$

Solving for $\mathbf{X}(i\,\omega)$ then gives

$$\mathbf{X}(i\,\omega) = ((i\,\omega)\,\mathbf{I} - \mathbf{F})^{-1}\,\mathbf{L}\,\mathbf{W}(i\,\omega), \tag{1.51}$$

Comparing to Equation (1.36) now reveals that actually we have

$$\mathcal{F}^{-1}\left[((i\,\omega)\,\mathbf{I} - \mathbf{F})^{-1}\right] = \exp(\mathbf{F}\,t)\,u(t), \tag{1.52}$$

where $u(t)$ is the Heaviside step function. This identity also provides one useful way to compute matrix exponentials.

Example 1.2 (Matrix exponential via Fourier transform)**.** *The matrix exponential considered in Example 1.1 can also be computed as*

$$\exp\left(\begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix} t\right) = \mathcal{F}^{-1}\left[\left(\begin{pmatrix}(i\,\omega) & 0\\ 0 & (i\,\omega)\end{pmatrix} - \begin{pmatrix}0 & 1\\ 0 & 0\end{pmatrix}\right)^{-1}\right] = \begin{pmatrix}1 & t\\ 0 & 1\end{pmatrix}. \tag{1.53}$$

1.5 Laplace transforms

Another often encountered method for solving linear time invariant differential equations is the *Laplace transform* (see, *e.g.*, Kreyszig, 1993). The Laplace transform of a function $f(t)$ (defined for all $t \geq 0$) is defined as

$$F(s) = \mathcal{L}[f(t)](s) = \int_0^\infty f(t)\exp(-st)\,\mathrm{d}t \tag{1.54}$$

and the inverse transform $f(t) = \mathcal{L}^{-1}[F(s)](t)$.

Just as for the Fourier transform, the usefulness of the Laplace transform comes from its property of reducing several often encountered "hard" differential equations into a "simple" subsidiary form which can be solved by algebraic manipulations. By inverse transforming the solution of the subsidiary equation, the solution to the original problem can be retrieved. Where a Fourier transform expresses the a function as a superposition of sinusoids, the Laplace transform expresses a function as a superposition of moments.

1.6 Numerical solutions of differential equations

For a generic non-linear differential equations of the form

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}(t), t), \quad \mathbf{x}(t_0) = \text{given}, \tag{1.55}$$

there is no general way to find an analytic solution. However, it is possible to approximate the solution numerically.

If we integrate the equation from t to $t + \Delta t$ we get

$$\mathbf{x}(t + \Delta t) = \mathbf{x}(t) + \int_t^{t+\Delta t} \mathbf{f}(\mathbf{x}(\tau), \tau)\,\mathrm{d}\tau. \tag{1.56}$$

If we knew how to compute the integral on the right hand side, we could generate the solution at time steps t_0, $t_1 = t_0 + \Delta t$, $t_2 = t_0 + 2\Delta$ iterating the above

equation:

$$\begin{aligned}
\mathbf{x}(t_0 + \Delta t) &= \mathbf{x}(t_0) + \int_{t_0}^{t_0+\Delta t} \mathbf{f}(\mathbf{x}(\tau), \tau)\, \mathrm{d}\tau \\
\mathbf{x}(t_0 + 2\Delta t) &= \mathbf{x}(t_0 + \Delta t) + \int_{t_0+\Delta t}^{t+2\Delta t} \mathbf{f}(\mathbf{x}(\tau), \tau)\, \mathrm{d}\tau \\
\mathbf{x}(t_0 + 3\Delta t) &= \mathbf{x}(t_0 + 2\Delta t) + \int_{t_0+2\Delta t}^{t+3\Delta t} \mathbf{f}(\mathbf{x}(\tau), \tau)\, \mathrm{d}\tau \\
&\vdots
\end{aligned} \tag{1.57}$$

It is now possible to derive various numerical methods by constructing approximations to the integrals on the right hand side. In the Euler method we use the approximation

$$\int_t^{t+\Delta t} \mathbf{f}(\mathbf{x}(\tau), \tau)\, \mathrm{d}\tau \approx \mathbf{f}(\mathbf{x}(t), t)\, \Delta t. \tag{1.58}$$

which leads to the following:

Algorithm 1.1 (Euler's method)**.** *Start from* $\hat{\mathbf{x}}(t_0) = \mathbf{x}(t_0)$ *and divide the integration interval* $[t_0, t]$ *into* n *steps* $t_0 < t_1 < t_2 < \ldots < t_n = t$ *such that* $\Delta t = t_{k+1} - t_k$*. At each step* k *approximate the solution as follows:*

$$\hat{\mathbf{x}}(t_{k+1}) = \hat{\mathbf{x}}(t_k) + \mathbf{f}(\hat{\mathbf{x}}(t_k), t_k)\, \Delta t. \tag{1.59}$$

The (global) order of a numerical integration methods can be defined to be the smallest exponent p such that if we numerically solve an ODE using $n = 1/\Delta t$ steps of length Δt, then there exists a constant K such that

$$|\hat{\mathbf{x}}(t_n) - \mathbf{x}(t_n)| \leq K\, \Delta t^p, \tag{1.60}$$

where $\hat{\mathbf{x}}(t_n)$ is the approximation and $\mathbf{x}(t_n)$ is the true solution. Because in Euler method, the first discarded term is of order Δt^2, the error of integrating over $1/\Delta t$ steps is proportional to Δt. Thus Euler method has order $p = 1$.

We can also improve the approximation by using trapezoidal approximation

$$\int_t^{t+\Delta t} \mathbf{f}(\mathbf{x}(\tau), \tau)\, \mathrm{d}\tau \approx \frac{\Delta t}{2} \left[\mathbf{f}(\mathbf{x}(t), t) + \mathbf{f}(\mathbf{x}(t + \Delta t), t + \Delta t)\right]. \tag{1.61}$$

which would lead to the approximate integration rule

$$\mathbf{x}(t_{k+1}) \approx \mathbf{x}(t_k) + \frac{\Delta t}{2} \left[\mathbf{f}(\mathbf{x}(t_k), t_k) + \mathbf{f}(\mathbf{x}(t_{k+1}), t_{k+1})\right]. \tag{1.62}$$

which is *implicit* rule in the sense that $\mathbf{x}(t_{k+1})$ appears also on the right hand side. To actually use such implicit rule, we would need to solve a non-linear equation

at each integration step which tends to be computationally too intensive when the dimensionality of $\mathbf{x}$ is high. Thus here we consider *explicit* rules only, where the next value $\mathbf{x}(t_{k+1})$ does not appear on the right hand side. If we now replace the term on the right hand side with its Euler approximation, we get the following *Heun's method*.

Algorithm 1.2 (Heun's method)**.** *Start from* $\hat{\mathbf{x}}(t_0) = \mathbf{x}(t_0)$ *and divide the integration interval* $[t_0, t]$ *into n steps* $t_0 < t_1 < t_2 < \ldots < t_n = t$ *such that* $\Delta t = t_{k+1} - t_k$*. At each step k approximate the solution as follows:*

$$\begin{aligned} \tilde{\mathbf{x}}(t_{k+1}) &= \hat{\mathbf{x}}(t_k) + \mathbf{f}(\hat{\mathbf{x}}(t_k), t_k)\,\Delta t \\ \hat{\mathbf{x}}(t_{k+1}) &= \hat{\mathbf{x}}(t_k) + \frac{\Delta t}{2}\left[\mathbf{f}(\hat{\mathbf{x}}(t_k), t_k) + \mathbf{f}(\tilde{\mathbf{x}}(t_{k+1}), t_{k+1})\right]. \end{aligned} \tag{1.63}$$

It can be shown that Heun's method has global order $p = 2$.

Another useful class of methods are the *Runge–Kutta methods*. The *classical 4th order Runge–Kutta method* is the following.

Algorithm 1.3 (4th order Runge–Kutta method)**.** *Start from* $\hat{\mathbf{x}}(t_0) = \mathbf{x}(t_0)$ *and divide the integration interval* $[t_0, t]$ *into n steps* $t_0 < t_1 < t_2 < \ldots < t_n = t$ *such that* $\Delta t = t_{k+1} - t_k$*. At each step k approximate the solution as follows:*

$$\begin{aligned} \Delta\mathbf{x}_k^1 &= \mathbf{f}(\hat{\mathbf{x}}(t_k), t_k)\,\Delta t \\ \Delta\mathbf{x}_k^2 &= \mathbf{f}(\hat{\mathbf{x}}(t_k) + \Delta\mathbf{x}_k^1/2, t_k + \Delta t/2)\,\Delta t \\ \Delta\mathbf{x}_k^3 &= \mathbf{f}(\hat{\mathbf{x}}(t_k) + \Delta\mathbf{x}_k^2/2, t_k + \Delta t/2)\,\Delta t \\ \Delta\mathbf{x}_k^4 &= \mathbf{f}(\hat{\mathbf{x}}(t_k) + \Delta\mathbf{x}_k^3, t_k + \Delta t)\,\Delta t \\ \hat{\mathbf{x}}(t_{k+1}) &= \hat{\mathbf{x}}(t_k) + \frac{1}{6}(\Delta\mathbf{x}_k^1 + 2\Delta\mathbf{x}_k^2 + 2\Delta\mathbf{x}_k^3 + \Delta\mathbf{x}_k^4). \end{aligned} \tag{1.64}$$

The above Runge–Kutta method can be derived by writing down the Taylor series expansion for the solution and by selecting coefficient such that many of the lower order terms cancel out. The order of this method is $p = 4$.

In fact, all the above integration methods are based on the Taylor series expansions of the solution. This is slightly problematic, because what happens in the case of SDEs is that the Taylor series expansion does not exists and all of the methods need to be modified at least to some extent. However, it is possible to replace Taylor series with so called Itô–Taylor series and then work out the analogous algorithms. The resulting algorithms are more complicated than the deterministic counterparts, because Itô–Taylor series is considerably more complicated than Taylor series. But we shall come back to this issue in Chapter 5.

There exists a wide class of other numerical ODE solvers as well. For example, all the above mentioned methods have a fixed step length, but there exists various variable step size methods which automatically adapt the step size. However, constructing variable step size methods for stochastic differential equations is much more involved than for deterministic equations and thus we shall not consider them here.

1.7 Picard–Lindelöf theorem

One important issue in differential equations is the question if the solution exists and whether it is unique. To analyze this questions, let's consider a generic equation of the form

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}(t), t), \quad \mathbf{x}(t_0) = \mathbf{x}_0, \tag{1.65}$$

where $\mathbf{f}(\mathbf{x}, t)$ is some given function. If the function $t \mapsto \mathbf{f}(\mathbf{x}(t), t)$ happens to be Riemann integrable, then we can integrate both sides from t_0 to t to yield

$$\mathbf{x}(t) = \mathbf{x}_0 + \int_{t_0}^{t} \mathbf{f}(\mathbf{x}(\tau), \tau)\, \mathrm{d}\tau. \tag{1.66}$$

We can now use this identity to find an approximate solution to the differential equation by the following *Picard's iteration* (see, *e.g.*, Tenenbaum and Pollard, 1985).

Algorithm 1.4 (Picard's iteration)**.** *Start from the initial guess* $\boldsymbol{\varphi}_0(t) = \mathbf{x}_0$*. Then compute approximations* $\boldsymbol{\varphi}_1(t), \boldsymbol{\varphi}_2(t), \boldsymbol{\varphi}_3(t), \ldots$ *via the following recursion:*

$$\boldsymbol{\varphi}_{n+1}(t) = \mathbf{x}_0 + \int_{t_0}^{t} \mathbf{f}(\boldsymbol{\varphi}_n(\tau), \tau)\, \mathrm{d}\tau \tag{1.67}$$

The above iteration, which we already used for finding the solution to linear differential equations in Section 1.2, can be shown to converge to the unique solution

$$\lim_{n\to\infty} \boldsymbol{\varphi}_n(t) = \mathbf{x}(t), \tag{1.68}$$

provided that $\mathbf{f}(\mathbf{x}, t)$ is continuous in both arguments and Lipschitz continuous in the first argument.

The implication of the above is the *Picard–Lindelöf theorem*, which says that under the above continuity conditions the differential equation has a solution and it is unique at a certain interval around $t = t_0$. We emphasis the innocent looking but important issue in the theorem: the function $\mathbf{f}(\mathbf{x}, t)$ needs to be *continuous*. This important, because in the case of stochastic differential equations the corresponding function will be *discontinuous everywhere* and thus we need a completely new existence theory for them.

Chapter 2

Pragmatic introduction to stochastic differential equations

2.1 Stochastic processes in physics, engineering, and other fields

The history of stochastic differential equations (SDEs) can be seen to have started form the classic paper of Einstein (1905), where he presented a mathematical connection between microscopic random motion of particles and the macroscopic diffusion equation. This is one of the results that proved the existence of the atom. Einstein's reasoning was roughly the following.

Example 2.1 (Microscopic motion of Brownian particles)**.** *Let τ be a small time interval and consider n particles suspended in liquid. During the time interval τ the x-coordinates of the particles will change by displacement Δ. The number of particles with displacement between Δ and $\Delta + \mathrm{d}\Delta$ is then*

$$\mathrm{d}n = n\,\phi(\Delta)\,\mathrm{d}\Delta, \tag{2.1}$$

where $\phi(\Delta)$ is the probability density of Δ, which can be assume to be symmetric $\phi(\Delta) = \phi(-\Delta)$ and differ from zero only for very small values of Δ.

Let $u(x,t)$ be the number of particles per unit volume. Then the number of particles at time $t + \tau$ located at $x + \mathrm{d}x$ is given as

$$u(x, t+\tau)\,\mathrm{d}x = \int_{-\infty}^{\infty} u(x+\Delta, t)\,\phi(\Delta)\,\mathrm{d}\Delta\,\mathrm{d}x. \tag{2.2}$$

Because τ is very small, we can put

$$u(x, t+\tau) = u(x,t) + \tau\,\frac{\partial u(x,t)}{\partial t}. \tag{2.3}$$

We can expand $u(x+\Delta, t)$ in powers of Δ:

$$u(x+\Delta, t) = u(x,t) + \Delta\,\frac{\partial u(x,t)}{\partial x} + \frac{\Delta^2}{2}\,\frac{\partial^2 u(x,t)}{\partial x^2} + \ldots \tag{2.4}$$

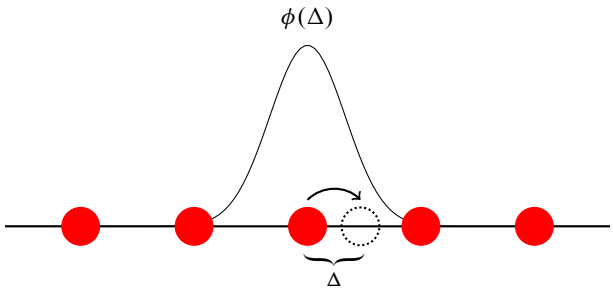


Figure 2.1: Illustration of Einstein's model of Brownian motion.

Substituting into (2.3) *and* (2.4) *into* (2.2) *gives*

$$u(x,t) + \tau \frac{\partial u(x,t)}{\partial t} = u(x,t) \int_{-\infty}^{\infty} \phi(\Delta)\, \mathrm{d}\Delta + \frac{\partial u(x,t)}{\partial x} \int_{-\infty}^{\infty} \Delta\, \phi(\Delta)\, \mathrm{d}\Delta + \frac{\partial^2 u(x,t)}{\partial x^2} \int_{-\infty}^{\infty} \frac{\Delta^2}{2} \phi(\Delta)\, \mathrm{d}\Delta + \ldots \quad (2.5)$$

where all the odd order terms vanish. If we recall that $\int_{-\infty}^{\infty} \phi(\Delta)\, \mathrm{d}\Delta = 1$ *and we put*

$$\int_{-\infty}^{\infty} \frac{\Delta^2}{2} \phi(\Delta)\, \mathrm{d}\Delta = D, \quad (2.6)$$

we get the diffusion equation

$$\frac{\partial u(x,t)}{\partial t} = D \frac{\partial^2 u(x,t)}{\partial x^2}. \quad (2.7)$$

This connection was significant during the time, because diffusion equation was only known as a macroscopic equation. Einstein was also able to derive a formula for D *in terms of microscopic quantities. From this, Einstein was able to compute the prediction for mean squared displacement of the particles as function of time:*

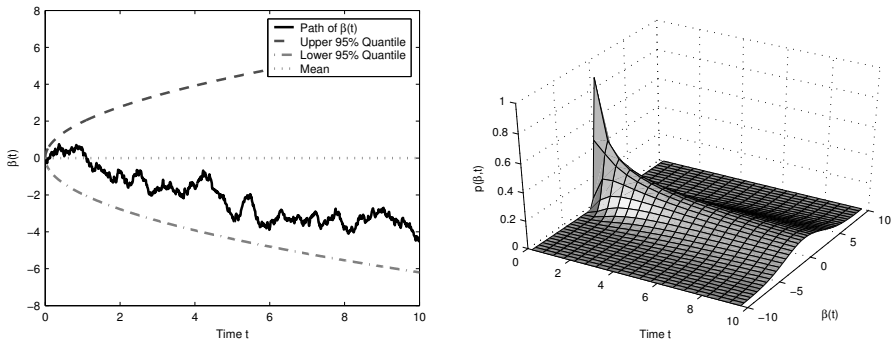
$$z(t) = \frac{R\,T}{N} \frac{1}{3\,\pi\,\eta\,r}\, t, \quad (2.8)$$

where η *is the viscosity of liquid,* r *is the diameter of the particles,* T *is the temperature,* R *is the gas constant, and* N *is the Avogadro constant.*

In modern terms, Brownian motion[1] (see Fig. 2.2) is an abstraction of a random walk process which has the property that each increment of it is independent. That is, direction and magnitude of each change of the process is completely random and independent of the previous changes. One way to think about Brownian motion is that it is the solution to the following stochastic differential equation

$$\frac{\mathrm{d}\beta(t)}{\mathrm{d}t} = w(t), \quad (2.9)$$

[1] In mathematics Brownian motion is also often called the Wiener process.



(a) A sample path **(b)** Probability density

Figure 2.2: Two views of Brownian motion: **(a)** a sample path and 95% quantiles, and **(b)** evolution of the probability density.

where $w(t)$ is a white random process. The term *white* here means that each the values $w(t)$ and $w(t')$ are independent whenever $t \neq t'$. We will later see that the probability density of the solution of this equation will solve the diffusion equation. However, at Einstein's time the theory of stochastic differential equations did not exists and therefore the reasoning was completely different.

A couple of years after Einstein's contribution Langevin (1908) presented an alternative construction of Brownian motion which leads to the same macroscopic properties. The reasoning in the article, which is outlined in the following, was more mechanical than in Einstein's derivation.

Example 2.2 (Langevin's model of Brownian motion)**.** *Consider a small particle suspended in liquid. Assume that there are two kinds of forces acting on the particle:*

1. *Friction force F_f, which by the Stokes law has the form:*

$$F_f = -6\,\pi\,\eta\,r\,v, \tag{2.10}$$

   where η is the viscosity, r is the diameter of particle and v is its velocity.

2. *Random force F_r caused by random collisions of the particles.*

Newton's law then gives

$$m\,\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = -6\,\pi\,\eta\,r\,\frac{\mathrm{d}x}{\mathrm{d}t} + F_r, \tag{2.11}$$

where m is the mass of the particle. Recall that

$$\begin{aligned} \frac{1}{2}\,\frac{\mathrm{d}(x^2)}{\mathrm{d}t} &= \frac{\mathrm{d}x}{\mathrm{d}t}\,x \\ \frac{1}{2}\,\frac{\mathrm{d}^2(x^2)}{\mathrm{d}t^2} &= \frac{\mathrm{d}^2x}{\mathrm{d}t^2}\,x + \left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2. \end{aligned} \tag{2.12}$$

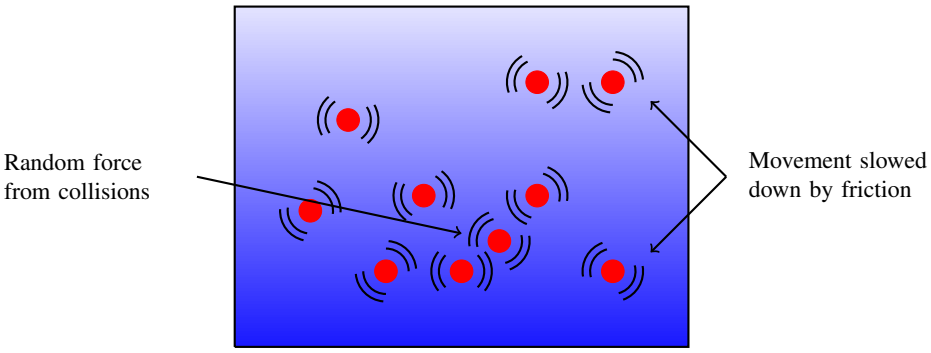


Figure 2.3: Illustration of Langevin's model of Brownian motion.

Thus if we multiply Equation (2.11) *with* x*, substitute the above identities, and take expectation we get*

$$\frac{m}{2}\,\mathrm{E}\left[\frac{\mathrm{d}^2(x^2)}{\mathrm{d}t^2}\right] - m\,\mathrm{E}\left[\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2\right] = -3\,\pi\,\eta\,r\,\mathrm{E}\left[\frac{\mathrm{d}(x^2)}{\mathrm{d}t}\right] + \mathrm{E}[F_r\,x]. \tag{2.13}$$

From statistical physics we know the relationship between the average kinetic energy and temperature:

$$m\,\mathrm{E}\left[\left(\frac{\mathrm{d}x}{\mathrm{d}t}\right)^2\right] = \frac{R\,T}{N}. \tag{2.14}$$

If we then assume that the random force and the position are uncorrelated, $\mathrm{E}[F_r\,x] = 0$ *and define a new variable* $\dot{z} = \mathrm{d}\,\mathrm{E}[x^2]/\mathrm{d}t$ *we get the differential equation*

$$\frac{m}{2}\frac{\mathrm{d}\dot{z}}{\mathrm{d}t} - \frac{R\,T}{N} = -3\,\pi\,\eta\,r\,\dot{z} \tag{2.15}$$

which has the general solution

$$\dot{z}(t) = \frac{R\,T}{N}\,\frac{1}{3\,\pi\,\eta\,r}\left[1 - \exp\left(\frac{6\,\pi\,\eta\,r}{m}\,t\right)\right]. \tag{2.16}$$

The exponential above goes to zero very quickly and thus the resulting mean squared displacement is nominally just the resulting constant multiplied with time:

$$z(t) = \frac{R\,T}{N}\,\frac{1}{3\,\pi\,\eta\,r}\,t, \tag{2.17}$$

which is exactly the same what Einstein obtained.

In the above model, Brownian motion is not actually seen as a solution to the white noise driven differential equation

$$\frac{\mathrm{d}\beta(t)}{\mathrm{d}t} = w(t), \tag{2.18}$$

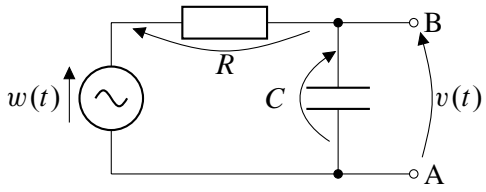


Figure 2.4: Example RC-circuit.

but instead, as the solution to equation of the form

$$\frac{\mathrm{d}^2\tilde{\beta}(t)}{\mathrm{d}t^2} = -c\,\frac{\mathrm{d}\tilde{\beta}(t)}{\mathrm{d}t} + w(t) \tag{2.19}$$

in the limit of vanishing time constant. The latter (Langevin's version) is sometimes called the physical Brownian motion and the former (Einstein's version) the mathematical Brownian motion. In these notes the term Brownian motion always means the mathematical Brownian motion.

Stochastic differential equations also arise other contexts. For example, the effect of thermal noise in electrical circuits and various kind of disturbances in telecommunications systems can be modeled as SDEs. In the following we present two such examples.

Example 2.3 (RC Circuit)**.** *Consider the simple RC circuit shown in Figure 2.4. In Laplace domain, the output voltage $V(s)$ can be expressed in terms of the input voltage $W(s)$ as follows:*

$$V(s) = \frac{1}{1 + R\,C\,s}\,W(s). \tag{2.20}$$

Inverse Laplace transform then gives the differential equation

$$\frac{\mathrm{d}v(t)}{\mathrm{d}t} = -\frac{1}{R\,C}\,v(t) + \frac{1}{R\,C}\,w(t). \tag{2.21}$$

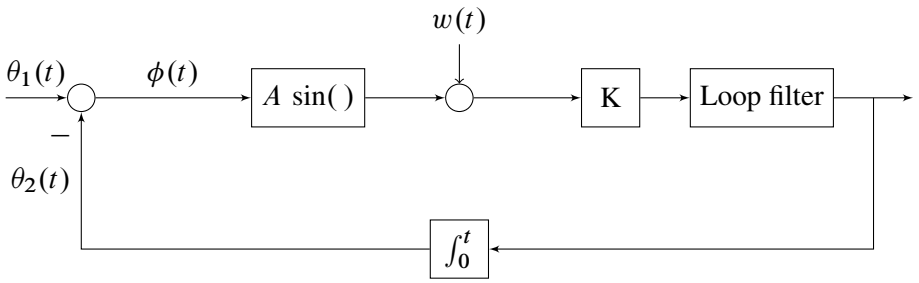


Figure 2.5: Simple phase locked loop (PLL) model.

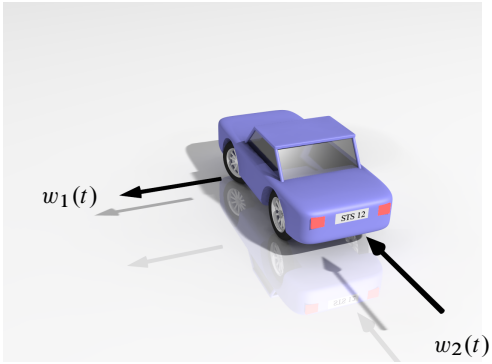


Figure 2.6: Illustration of a car's dynamic model.

For the purposes of studying the response of the circuit to noise, we can now replace the input voltage with a white noise process $w(t)$ and analyze the properties of the resulting equation.

Example 2.4 (Phase locked loop (PLL))**.** *Phase locked loops (PLLs) are telecommunications system devices, which can be used to automatically synchronize a demodulator with a carrier signal. A simple mathematical model of PLL is shown in Figure 2.5 (see, Viterbi, 1966), where $w(t)$ models disturbances (noise) in the system. In the case that there is no loop filter at all and when the input is a constant-frequency sinusoid $\theta_1(t) = (\omega - \omega_0)\, t + \theta$, the differential equation for the system becomes*

$$\frac{\mathrm{d}\phi}{\mathrm{d}t} = (\omega - \omega_0) - A\ K\ \sin\phi(t) - K\, w(t). \tag{2.22}$$

It is now possible to analyze the properties of PLL in the presence of noise by analyzing the properties of this stochastic differential equation (Viterbi, 1966).

Stochastic differential equations can also be used for modeling dynamic phenomena, where the exact dynamics of the system are uncertain. For example, the motion model of a car cannot be exactly written down if we do not know all the external forces affecting the car and the acts of the driver. However, the unknown sub-phenomena can be modeled as stochastic processes, which leads to stochastic differential equations. This kind of modeling principle of representing uncertainties as random variables is sometimes called *Bayesian modeling*. Stochastic differential equation models of this kind and commonly used in navigation and control systems (see, *e.g.*, Jazwinski, 1970; Bar-Shalom et al., 2001; Grewal and Andrews, 2001). Stock prices can also be modeled using stochastic differential equations and this kind of models are indeed commonly used in analysis and pricing of stocks and related quantities (Øksendal, 2003).

Example 2.5 (Dynamic model of a car)**.** *The dynamics of the car in 2d (x_1, x_2) are governed by the* Newton's law *(see Fig. 2.6):*

$$\mathbf{f}(t) = m\, \mathbf{a}(t), \tag{2.23}$$

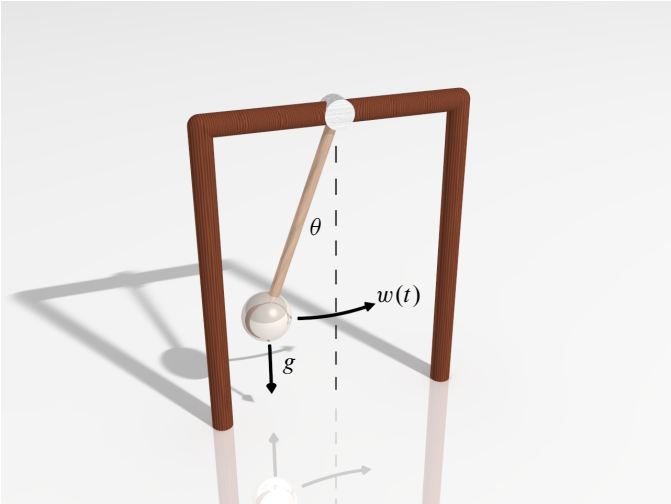


Figure 2.7: An illustration for the pendulum example.

where $\mathbf{a}(t)$ *is the acceleration,* m *is the mass of the car, and* $\mathbf{f}(t)$ *is a vector of (unknown) forces acting the car. Let's now model* $\mathbf{f}(t)/m$ *as a two-dimensional* white random process*:*

$$\begin{aligned} \frac{\mathrm{d}^2 x_1}{\mathrm{d}t^2} &= w_1(t), \\ \frac{\mathrm{d}^2 x_2}{\mathrm{d}t^2} &= w_2(t). \end{aligned} \tag{2.24}$$

If we define $x_3(t) = \mathrm{d}x_1/\mathrm{d}t$ *,* $x_4(t) = \mathrm{d}x_2/\mathrm{d}t$ *, then the model can be written as a first order* system of differential equations*:*

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}}_{\mathbf{F}} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} + \underbrace{\begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}}_{\mathbf{L}} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}. \tag{2.25}$$

In shorter matrix form *this can be written as a linear differential equation model:*

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{F}\,\mathbf{x} + \mathbf{L}\,\mathbf{w}.$$

Example 2.6 (Noisy pendulum)**.** *The differential equation for a simple pendulum (see Fig. 2.7) with unit length and mass can be written as:*

$$\ddot{\theta} = -g\,\sin(\theta) + w(t), \tag{2.26}$$

where θ is the angle, g is the gravitational acceleration and $w(t)$ is a random noise process. This model can be converted converted into the following state space model:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} = \begin{pmatrix}\theta_2\\ -g\ \sin(\theta_1)\end{pmatrix} + \begin{pmatrix}0\\ 1\end{pmatrix} w(t). \tag{2.27}$$

This can be seen to be a special case of equations of the form

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}) + \mathbf{L}\,\mathbf{w}, \tag{2.28}$$

where $\mathbf{f}(\mathbf{x})$ is a non-linear function.

Example 2.7 (Black–Scholes model)**.** *In the Black–Scholes model the asset (e.g., stock price) x is assumed to follow geometric Brownian motion*

$$\mathrm{d}x = \mu\, x\ \mathrm{d}t + \sigma\, x\ \mathrm{d}\beta. \tag{2.29}$$

where $\mathrm{d}\beta$ is a Brownian motion increment, μ is a drift constant and σ is a volatility constant. If we formally divide by $\mathrm{d}t$, this equation can be heuristically interpreted as a differential equation

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \mu\, x + \sigma\, x\, w, \tag{2.30}$$

where $w(t)$ is a white random process. This equation is now an example of more general multiplicative noise models of the form

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}) + \mathbf{L}(\mathbf{x})\,\mathbf{w}. \tag{2.31}$$

2.2 Differential equations with driving white noise

As discussed in the previous section, many time-varying phenomena in various fields in science and engineering can be modeled as differential equations of the form

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}, t) + \mathbf{L}(\mathbf{x}, t)\,\mathbf{w}(t). \tag{2.32}$$

where $\mathbf{w}(t)$ is some vector of forcing functions.

We can think a *stochastic differential equation (SDE)* as an equation of the above form where the forcing function is a stochastic process. One motivation for studying such equations is that various physical phenomena can be modeled as random processes (*e.g.*, thermal motion) and when such a phenomenon enters a physical system, we get a model of the above SDE form. Another motivation is that in Bayesian statistical modeling *unknown* forces are naturally modeled as *random* forces which again leads to SDE type of models. Because the forcing function is random, the solution to the stochastic differential equation is a random process as well. With a different realization of the noise process we get a different solution. For this reason the particular solutions of the equations are not often of interest, but

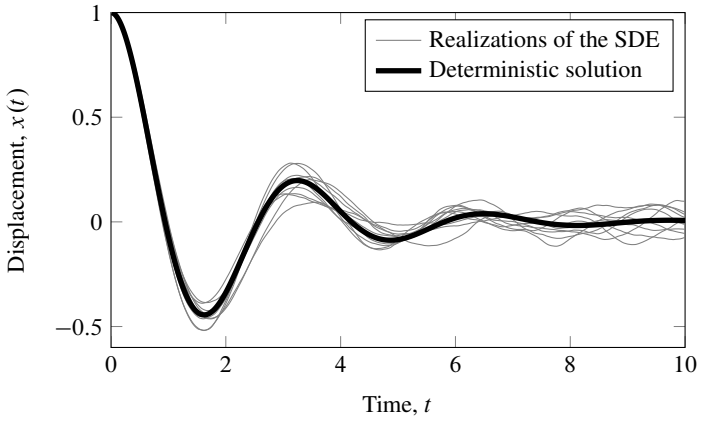


Figure 2.8: Solutions of the spring model in Equation (1.1) when the input is white noise. The solution of the SDE is different for each realization of noise process. We can also compute the mean of the solutions, which in the case of linear SDE corresponds to the deterministic solution with zero noise.

instead, we aim to determine the statistics of the solutions over all realizations. An example of SDE solution is given in Figure 2.8.

In the context of SDEs, the term $\mathbf{f}(\mathbf{x}, t)$ in Equation (2.32) is called the drift function which determines the nominal dynamics of the system, and $\mathbf{L}(\mathbf{x}, t)$ is the dispersion matrix which determines how the noise $\mathbf{w}(t)$ enters the system. This indeed is the most general form of SDE that we discuss in the document. Although it would be tempting to generalize these equations to $\mathrm{d}\mathbf{x}/\mathrm{d}t = \mathbf{f}(\mathbf{x}, \mathbf{w}, t)$, it is not possible in the present theory. We shall discuss the reason for this later in this document.

The unknown function usually modeled as Gaussian and "white" in the sense that $\mathbf{w}(t)$ and $\mathbf{w}(\tau)$ are uncorrelated (and independent) for all $t \neq s$. The term *white* arises from the property that the power spectrum (or actually, the spectral density) of white noise is constant (flat) over all frequencies. White light is another phenomenon which has this same property and hence the name.

In mathematical sense white noise process can be defined as follows:

Definition 2.1 (White noise)**.** *White noise process $\mathbf{w}(t) \in \mathbb{R}^s$ is a random function with the following properties:*

1. *$\mathbf{w}(t_1)$ and $\mathbf{w}(t_2)$ are independent if $t_1 \neq t_2$.*

2. *$t \mapsto \mathbf{w}(t)$ is a Gaussian process with zero mean and Dirac-delta-correlation:*

$$\begin{aligned} \mathbf{m}_w(t) &= \mathrm{E}[\mathbf{w}(t)] = \mathbf{0} \\ \mathbf{C}_w(t, s) &= \mathrm{E}[\mathbf{w}(t)\,\mathbf{w}^{\mathsf{T}}(s)] = \delta(t - s)\,\mathbf{Q}, \end{aligned} \tag{2.33}$$

   where $\mathbf{Q}$ is the spectral density of the process.

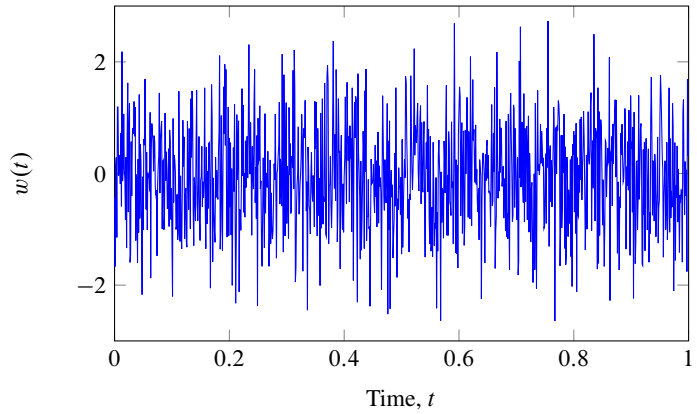


Figure 2.9: White noise.

From the above properties we can also deduce the following somewhat peculiar properties of white noise:

- The sample path $t \mapsto \mathbf{w}(t)$ is discontinuous almost everywhere.
- White noise is unbounded and it takes arbitrarily large positive and negative values at any finite interval.

An example of a scalar white noise process realization is shown in Figure 2.9.

It is also possible to use non-Gaussian driving functions in SDEs such as Poisson processes or more general Lévy processes (see, *e.g.*, Applebaum, 2004), but here we will always assume that the driving function is Gaussian.

2.3 Heuristic solutions of linear SDEs

Let's first consider linear time-invariant stochastic differential equations (LTI SDEs) of the form

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \mathbf{F}\,\mathbf{x}(t) + \mathbf{L}\,\mathbf{w}(t), \qquad \mathbf{x}(0) \sim \mathrm{N}(\mathbf{m}_0, \mathbf{P}_0), \tag{2.34}$$

where $\mathbf{F}$ and $\mathbf{L}$ are some constant matrices the white noise process $\mathbf{w}(t)$ has zero mean and a given spectral density $\mathbf{Q}$. Above, we have specified a random initial condition for the equation such that at initial time $t = 0$ the solutions should be Gaussian with a given mean $\mathbf{m}_0$ and covariance $\mathbf{P}_0$.

If we pretend for a while that the driving process $\mathbf{w}(t)$ is deterministic and continuous, we can form the general solution to the differential equation as follows:

$$\mathbf{x}(t) = \exp(\mathbf{F}\,t)\;\mathbf{x}(0) + \int_0^t \exp(\mathbf{F}\,(t-\tau))\;\mathbf{L}\,\mathbf{w}(\tau)\;\mathrm{d}\tau, \tag{2.35}$$

where $\exp(\mathbf{F}\,t)$ is the matrix exponential function.

We can now take a "leap of faith" and hope that this solutions is valid also when $\mathbf{w}(t)$ is a white noise process. It turns out that it indeed is, but just because the differential equation happens to be linear (we'll come back to this issue in next chapter). However, it is enough for our purposes for now. The solution also turns out to be Gaussian, because the noise process is Gaussian and a linear differential equation can be considered as a linear operator acting on the noise process (and the initial condition).

Because white noise process has zero mean, taking expectations from the both sides of Equation (2.35) gives

$$\mathrm{E}[\mathbf{x}(t)] = \exp(\mathbf{F}\,t)\;\mathbf{m}_0, \tag{2.36}$$

which is thus the expected value of the SDE solutions over all realizations of noise. The mean function is here denoted as $\mathbf{m}(t) = \mathrm{E}[\mathbf{x}(t)]$.

The covariance of the solution can be derived by substituting the solution into the definition of covariance and by using the delta-correlation property of white noise, which results in

$$\begin{aligned}&\mathrm{E}\left[(\mathbf{x}(t) - \mathbf{m}(t))\,(\mathbf{x}(t) - \mathbf{m})^\mathsf{T}\right] \\ &= \exp(\mathbf{F}\,t)\;\mathbf{P}_0\;\exp(\mathbf{F}\,t)^\mathsf{T} + \int_0^t \exp(\mathbf{F}\,(t-\tau))\;\mathbf{L}\,\mathbf{Q}\,\mathbf{L}^\mathsf{T}\,\exp(\mathbf{F}\,(t-\tau))^\mathsf{T}\,\mathrm{d}\tau.\end{aligned} \tag{2.37}$$

In this document we denote the covariance as $\mathbf{P}(t) = \mathrm{E}\left[(\mathbf{x}(t) - \mathbf{m}(t))\,(\mathbf{x}(t) - \mathbf{m})^\mathsf{T}\right]$.

By differentiating the mean and covariance solutions and collecting the terms we can also derive the following differential equations for the mean and covariance:

$$\begin{aligned}\frac{\mathrm{d}\mathbf{m}(t)}{\mathrm{d}t} &= \mathbf{F}\,\mathbf{m}(t) \\ \frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t} &= \mathbf{F}\,\mathbf{P}(t) + \mathbf{P}(t)\,\mathbf{F}^\mathsf{T} + \mathbf{L}\,\mathbf{Q}\,\mathbf{L}^\mathsf{T},\end{aligned} \tag{2.38}$$

Example 2.8 (Stochastic spring model)**.** *If in the spring model of Equation* (1.4)*, we replace the input force with a white noise with spectral density q, we get the following LTI SDE:*

$$\underbrace{\begin{pmatrix}\frac{\mathrm{d}x_1(t)}{\mathrm{d}t} \\ \frac{\mathrm{d}x_2(t)}{\mathrm{d}t}\end{pmatrix}}_{\mathrm{d}\mathbf{x}(t)/\mathrm{d}t} = \underbrace{\begin{pmatrix}0 & 1 \\ -\nu^2 & -\gamma\end{pmatrix}}_{\mathbf{F}}\underbrace{\begin{pmatrix}x_1(t) \\ x_2(t)\end{pmatrix}}_{\mathbf{x}} + \underbrace{\begin{pmatrix}0 \\ 1\end{pmatrix}}_{\mathbf{L}}\,w(t). \tag{2.39}$$

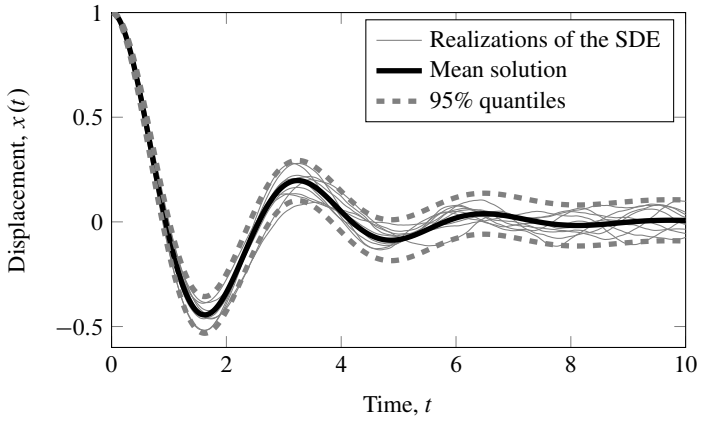


Figure 2.10: Solutions, theoretical mean, and the 95% quantiles for the spring model in Equation (1.1) when the input is white noise.

The equations for the mean and covariance are thus given as

$$\begin{aligned}
\begin{pmatrix} \frac{\mathrm{d}m_1}{\mathrm{d}t} \\ \frac{\mathrm{d}m_2}{\mathrm{d}t} \end{pmatrix} &= \begin{pmatrix} 0 & 1 \\ -\nu^2 & -\gamma \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} \\
\begin{pmatrix} \frac{\mathrm{d}P_{11}}{\mathrm{d}t} & \frac{\mathrm{d}P_{12}}{\mathrm{d}t} \\ \frac{\mathrm{d}P_{21}}{\mathrm{d}t} & \frac{\mathrm{d}P_{22}}{\mathrm{d}t} \end{pmatrix} &= \begin{pmatrix} 0 & 1 \\ -\nu^2 & -\gamma \end{pmatrix} \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix} \\
&\quad + \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -\nu^2 & -\gamma \end{pmatrix}^{\mathsf{T}} + \begin{pmatrix} 0 & 0 \\ 0 & q \end{pmatrix}
\end{aligned} \tag{2.40}$$

Figure 2.10 shows the theoretical mean and the 95% quantiles computed from the variances $P_{11}(t)$ along with trajectories from the stochastic spring model.

Despite the heuristic derivation, Equations (2.38) are indeed the correct differential equations for the mean and covariance. But it is easy to demonstrate that one has to be extremely careful in extrapolation of deterministic differential equation results to stochastic setting.

Note that we can indeed derive the first of the above equations simply by taking the expectations from both sides of Equation (2.34):

$$\mathrm{E}\left[\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t}\right] = \mathrm{E}\left[\mathbf{F}\,\mathbf{x}(t)\right] + \mathrm{E}\left[\mathbf{L}\,\mathbf{w}(t)\right], \tag{2.41}$$

Exchanging the order of expectation and differentiation, using the linearity of expectation and recalling that white noise has zero mean then results in correct mean differential equation. We can now attempt to do the same for the covariance. By the chain rule of ordinary calculus we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[(\mathbf{x}-\mathbf{m})\;(\mathbf{x}-\mathbf{m})^{\mathsf{T}}\right] = \left(\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} - \frac{\mathrm{d}\mathbf{m}}{\mathrm{d}t}\right)(\mathbf{x}-\mathbf{m})^{\mathsf{T}} + (\mathbf{x}-\mathbf{m})\left(\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} - \frac{\mathrm{d}\mathbf{m}}{\mathrm{d}t}\right)^{\mathsf{T}}, \tag{2.42}$$

Substituting the time derivatives to the right hand side and taking expectation then results in

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\,\mathrm{E}\Big[(\mathbf{x}-\mathbf{m})\;(\mathbf{x}-\mathbf{m})^{\mathsf{T}}\Big] = \mathbf{F}\;\mathrm{E}\Big[(\mathbf{x}(t)-\mathbf{m}(t))\;(\mathbf{x}(t)-\mathbf{m}(t))^{\mathsf{T}}\Big] \\ + \mathrm{E}\Big[(\mathbf{x}(t)-\mathbf{m}(t))\;(\mathbf{x}(t)-\mathbf{m}(t))^{\mathsf{T}}\Big]\;\mathbf{F}^{\mathsf{T}},\end{aligned} \tag{2.43}$$

which implies the covariance differential equation

$$\frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t} = \mathbf{F}\,\mathbf{P}(t) + \mathbf{P}(t)\,\mathbf{F}^{\mathsf{T}}. \tag{2.44}$$

But this equation is *wrong*, because the term $\mathbf{L}(t)\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}(t)$ is missing from the right hand side. Our mistake was to assume that we can use the product rule in Equation (2.42), but in fact we cannot. This is one of the peculiar features of stochastic calculus and it is also a warning sign that we should not take our "leap of faith" too far when analyzing solutions of SDEs via formal extensions of deterministic ODE solutions.

2.4 Heuristic solutions of non-linear SDEs

We could now attempt to analyze differential equations of the form

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{f}(\mathbf{x},t) + \mathbf{L}(\mathbf{x},t)\,\mathbf{w}(t), \tag{2.45}$$

where $\mathbf{f}(\mathbf{x},t)$ and $\mathbf{L}(\mathbf{x},t)$ are non-linear functions and $\mathbf{w}(t)$ is a white noise process with a spectral density $\mathbf{Q}$. Unfortunately, we cannot take the same kind of "leap of faith" from deterministic solutions as in the case of linear differential equations, because we could not solve even the deterministic differential equation.

An attempt to generalize the numerical methods for deterministic differential equations discussed in previous chapter will fail as well, because the basic requirement in almost all of those methods is continuity of the right hand side, and in fact, even differentiability of several orders. Because white noise is discontinuous everywhere, the right hand side is discontinuous everywhere, and is certainly not differentiable anywhere either. Thus we are in trouble.

We can, however, generalize the Euler method (leading to Euler–Maruyama method) to the present stochastic setting, because it does not explicitly require continuity. From that, we get an iteration of the form

$$\hat{\mathbf{x}}(t_{k+1}) = \hat{\mathbf{x}}(t_k) + \mathbf{f}(\hat{\mathbf{x}}(t_k),t_k)\,\Delta t + \mathbf{L}(\hat{\mathbf{x}}(t_k),t_k)\,\Delta\beta_k, \tag{2.46}$$

where $\Delta\beta_k$ is a Gaussian random variable with distribution $\mathrm{N}(\mathbf{0},\mathbf{Q}\,\Delta t)$. Note that it is indeed the variance which is proportional to Δt, not the standard derivation as we might expect. This results from the peculiar properties of stochastic differential equations. Anyway, we can use the above method to simulate trajectories from

stochastic differential equations and the result converges to the true solution in the limit $\Delta t \to 0$. However, the convergence is quite slow as the order of convergence is only $p = 1/2$.

In the case of SDEs, the convergence order definition is a bit more complicated, because we can talk about path-wise approximations, which corresponds to approximating the solution with fixed $\mathbf{w}(t)$. These are also called strong solution and give rise to strong order of convergence. But we can also think of approximating the probability density or the moments of the solutions. These give rise to weak solutions and weak order of convergence. We will come back to these later.

2.5 The problem of solution existence and uniqueness

Let's now attempt to analyze the uniqueness and existence of the equation

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}, t) + \mathbf{L}(\mathbf{x}, t)\,\mathbf{w}(t), \tag{2.47}$$

using the Picard–Lindelöf theorem presented in the previous chapter. The basic assumption in the theorem for the right hand side of the differential equation were:

- Continuity in both arguments.
- Lipschitz continuity in the first argument.

Unfortunately, the first of these fails miserably, because white noise is discontinuous everywhere. However, a small blink of hope is implied by that $\mathbf{f}(\mathbf{x}, t)$ might indeed be Lipschitz continuous in the first argument, as well as $\mathbf{L}(\mathbf{x}, t)$. However, without extending the Pickard–Lindelöf theorem we cannot determine the existence or uniqueness of stochastic differential equations.

It turns out that a stochastic analog of Picard iteration will indeed lead to the solution to the existence and uniqueness problem also in the stochastic case. But before going into that we need to make the theory of stochastic differential equations mathematically meaningful.

Chapter 3

Itô calculus and stochastic differential equations

3.1 The stochastic integral of Itô

As discussed in the previous chapter, a stochastic differential equation can be heuristically considered as a vector differential equation of the form

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}, t) + \mathbf{L}(\mathbf{x}, t)\,\mathbf{w}(t), \tag{3.1}$$

where $\mathbf{w}(t)$ is a zero mean white Gaussian process. However, although this is sometimes true, it is not the whole truth. The aim in this section is to clarify what really is the logic behind stochastic differential equations and how we should treat them.

The problem in the above equation is that it cannot be a differential equation in traditional sense, because the ordinary theory of differential equations does not permit discontinuous functions such as $\mathbf{w}(t)$ in differential equations (recall the problem with the Picard–Lindelöf theorem). And the problem is not purely theoretical, because the solution actually turns out to depend on infinitesimally small differences in mathematical definitions of the noise and thus without further restrictions the solution would not be unique even with a given realization of white noise $\mathbf{w}(t)$.

Fortunately, there is a solution to this problem, but in order to find it we need to reduce the problem to definition of a new kind of integral called the *Itô integral*, which is an integral with respect to a stochastic process. In order to do that, let's first formally integrate the differential equation from some initial time t_0 to final time t:

$$\mathbf{x}(t) - \mathbf{x}(t_0) = \int_{t_0}^{t} \mathbf{f}(\mathbf{x}(t), t)\,\mathrm{d}t + \int_{t_0}^{t} \mathbf{L}(\mathbf{x}(t), t)\,\mathbf{w}(t)\,\mathrm{d}t. \tag{3.2}$$

The first integral on the right hand side is just a normal integral with respect to time and can be defined as a Riemann integral of $t \mapsto \mathbf{f}(\mathbf{x}(t), t)$ or as a Lebesgue integral with respect to the Lebesgue measure, if more generality is desired.

The second integral is the problematic one. First of all, it cannot be defined as Riemann integral due to the unboundedness and discontinuity of the white noise process. Recall that in the Riemannian sense the integral would be defined as the following kind of limit:

$$\int_{t_0}^{t} \mathbf{L}(\mathbf{x}(t), t)\, \mathbf{w}(t)\, \mathrm{d}t = \lim_{n\to\infty} \sum_{k} \mathbf{L}(\mathbf{x}(t_k^*), t_k^*)\, \mathbf{w}(t_k^*)\, (t_{k+1} - t_k), \tag{3.3}$$

where $t_0 < t_1 < \ldots < t_n = t$ and $t_k^* \in [t_k, t_{k+1}]$. In the context of Riemann integrals so called upper and lower sums are defined as the selections of t_k^* such that the integrand $\mathbf{L}(\mathbf{x}(t_k^*), t_k^*)\, \mathbf{w}(t_k^*)$ has its maximum and minimum values, respectively. The Riemann integral is defined if the upper and lower sums converge to the same value, which is then defined to be the value of the integral. In the case of white noise it happens that $\mathbf{w}(t_k^*)$ is not bounded and takes arbitrarily small and large values at every finite interval, and thus the Riemann integral does not converge.

We could also attempt to define it as a Stieltjes integral which is more general than the Riemann integral. For that definition we need to interpret the increment $\mathbf{w}(t)\, \mathrm{d}t$ as an increment of another process $\boldsymbol{\beta}(t)$ such that the integral becomes

$$\int_{t_0}^{t} \mathbf{L}(\mathbf{x}(t), t)\, \mathbf{w}(t)\, \mathrm{d}t = \int_{t_0}^{t} \mathbf{L}(\mathbf{x}(t), t)\, \mathrm{d}\boldsymbol{\beta}(t). \tag{3.4}$$

It turns out that a suitable process for this purpose is the Brownian motion which we already discussed in the previous chapter:

Definition 3.1 (Brownian motion)**.** *Brownian motion $\boldsymbol{\beta}(t)$ is a process with the following properties:*

1. *Any increment $\Delta\boldsymbol{\beta}_k = \boldsymbol{\beta}(t_{k+1}) - \boldsymbol{\beta}(t_k)$ is a zero mean Gaussian random variable with covariance $\mathbf{Q}\, \Delta t_k$, where $\mathbf{Q}$ is the diffusion matrix of the Brownian motion and $\Delta t_k = t_{k+1} - t_k$.*

2. *When the time spans of increments do not overlap, the increments are independent.*

Some further properties of Brownian motion which result from the above are the following:

1. Brownian motion $t \mapsto \boldsymbol{\beta}(t)$ has a discontinuous derivative everywhere.

2. White noise can be considered as the formal derivative of Brownian motion $\mathbf{w}(t) = \mathrm{d}\boldsymbol{\beta}(t)/\mathrm{d}t$.

An example of a scalar Brownian motion realization is shown in Figure 3.1.

Unfortunately, the definition of the latter integral in Equation (3.2) in terms of increments of Brownian motion as in Equation (3.4) does not solve our existence

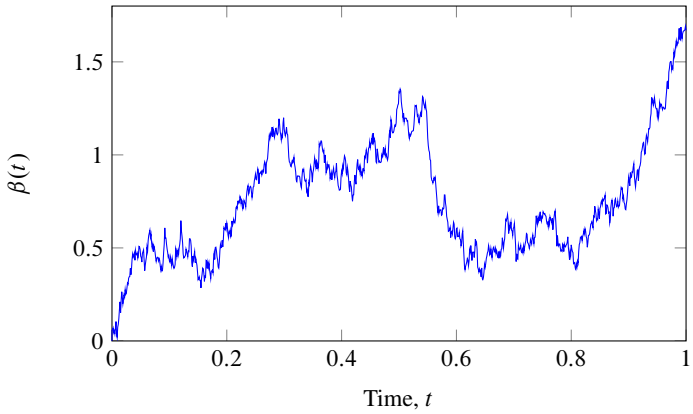


Figure 3.1: A realization trajectory of Brownian motion, where the derivative is discontinuous everywhere. White noise can be considered the formal derivative of Brownian motion.

problem. The problem is the everywhere discontinuous derivative of $\boldsymbol{\beta}(t)$ which makes it too irregular for the defining sum of the Stieltjes integral to converge. Unfortunately, the same happens with the Lebesgue integral. Recall that both Stieltjes and Lebesgue integrals are essentially defined as limits of the form

$$\int_{t_0}^{t} \mathbf{L}(\mathbf{x}(t), t)\, \mathrm{d}\boldsymbol{\beta} = \lim_{n\to\infty} \sum_k \mathbf{L}(\mathbf{x}(t_k^*), t_k^*)\, [\boldsymbol{\beta}(t_{k+1}) - \boldsymbol{\beta}(t_k)], \tag{3.5}$$

where $t_0 < t_1 < \ldots < t_n$ and $t_k^* \in [t_k, t_{k+1}]$. The core problem in both of these definitions is that they would require the limit to be independent of the position on the interval $t_k^* \in [t_k, t_{k+1}]$. But for integration with respect to Brownian motion this is not the case. Thus, the Stieltjes or Lebesgue integral definitions does not work either.

The solution to the problem is the Itô stochastic integral which is based on the observation that if we fix the choice to $t_k^* = t_k$, then the limit becomes unique. The Itô integral can thus be defined as the limit

$$\int_{t_0}^{t} \mathbf{L}(\mathbf{x}(t), t)\, \mathrm{d}\boldsymbol{\beta}(t) = \lim_{n\to\infty} \sum_k \mathbf{L}(\mathbf{x}(t_k), t_k)\, [\boldsymbol{\beta}(t_{k+1}) - \boldsymbol{\beta}(t_k)], \tag{3.6}$$

which is a sensible definition of the stochastic integral required for the SDE.

The stochastic differential equation (2.32) can now be defined to actually refer to the corresponding (Itô) integral equation

$$\mathbf{x}(t) - \mathbf{x}(t_0) = \int_{t_0}^{t} \mathbf{f}(\mathbf{x}(t), t)\, \mathrm{d}t + \int_{t_0}^{t} \mathbf{L}(\mathbf{x}(t), t)\, \mathrm{d}\boldsymbol{\beta}(t), \tag{3.7}$$

which should be true for arbitrary t_0 and t.

We can now take a step backwards and return from this stochastic integral equation to the differential equation as follows. If we choose the integration limits in Equation (3.7) to be t and $t + \mathrm{d}t$, where $\mathrm{d}t$ is "small", we can write the equation in the differential form

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\,\mathrm{d}t + \mathbf{L}(\mathbf{x}, t)\,\mathrm{d}\boldsymbol{\beta}, \tag{3.8}$$

which should be interpreted as shorthand for the integral equation. The above is the form which is most often used in literature on stochastic differential equations (*e.g.*, Øksendal, 2003; Karatzas and Shreve, 1991). We can now formally divide by $\mathrm{d}t$ to obtain a differential equation:

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}, t) + \mathbf{L}(\mathbf{x}, t)\,\frac{\mathrm{d}\boldsymbol{\beta}}{\mathrm{d}t}, \tag{3.9}$$

which shows that also here white noise can be interpreted as the formal derivative of Brownian motion. However, due to non-classical transformation properties of the Itô differentials, one has to be very careful in working with such formal manipulations.

It is now also easy to see why we are not permitted to consider more general differential equations of the form

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}(t), \mathbf{w}(t), t), \tag{3.10}$$

where the white noise $\mathbf{w}(t)$ enters the system through a non-linear transformation. There is no way to rewrite this equation as a stochastic integral with respect to a Brownian motion and thus we cannot define the mathematical meaning of this equation. More generally, white noise should not be thought of as an entity as such, but it only exists as the formal derivative of Brownian motion. Therefore only linear functions of white noise have a meaning whereas non-linear functions do not.

Let's now take a short excursion to how Itô integrals are often treated in stochastic analysis. In the above treatment we have only considered stochastic integration of the term $\mathbf{L}(\mathbf{x}(t), t)$, but the definition can be extended to arbitrary Itô processes $\boldsymbol{\Theta}(t)$, which are "adapted" to the Brownian motion $\boldsymbol{\beta}(t)$ to be integrated over. Being "adapted" means that $\boldsymbol{\beta}(t)$ is the only stochastic "driving force" in $\boldsymbol{\Theta}(t)$ in the same sense that $\mathbf{L}(\mathbf{x}(t), t)$ was generated as a function of $\mathbf{x}(t)$, which in turn is generated though the differential equation, where the only stochastic driver is the Brownian motion. This adaptation can also be denoted by including the "event space element" ω as an argument to the function $\boldsymbol{\Theta}(t, \omega)$ and Brownian motion $\boldsymbol{\beta}(t, \omega)$. The resulting Itô integral is then defined as the limit

$$\int_{t_0}^{t} \boldsymbol{\Theta}(t, \omega)\,\mathrm{d}\boldsymbol{\beta}(t, \omega) = \lim_{n\to\infty} \sum_{k} \boldsymbol{\Theta}(t_k, \omega)\,[\boldsymbol{\beta}(t_{k+1}, \omega) - \boldsymbol{\beta}(t_k, \omega)]. \tag{3.11}$$

Actually, the definition is slightly more complicated (see Karatzas and Shreve, 1991; Øksendal, 2003), but the basic principle is the above. Furthermore, if $\boldsymbol{\Theta}(t, \omega)$

is such an adapted process, then according to the martingale representation theorem it can always be represented as the solution to a suitable Itô stochastic differential equation. Malliavin calculus (Nualart, 2006) provides the tools for finding such an equation in practice. However, this kind of analysis would require us to use the full measure theoretical formulation of the Itô stochastic integral which we will not do here.

3.2 Itô formula

Consider the stochastic integral

$$\int_0^t \beta(t)\,\mathrm{d}\beta(t) \tag{3.12}$$

where $\beta(t)$ is a standard Brownian motion, that is, scalar Brownian motion with diffusion matrix $Q = 1$. Based on ordinary calculus we would expect the value of this integral to be $\beta^2(t)/2$, but it is the wrong answer. If we select a partition $0 = t_0 < t_1 < \ldots < t_n = t$, we get by rearranging the terms

$$\begin{aligned}\int_0^t \beta(t)\,\mathrm{d}\beta(t) &= \lim_{n\to\infty} \sum_k \beta(t_k)[\beta(t_{k+1}) - \beta(t_k)] \\ &= \lim_{n\to\infty} \sum_k \left[-\frac{1}{2}(\beta(t_{k+1}) - \beta(t_k))^2 + \frac{1}{2}(\beta^2(t_{k+1}) - \beta^2(t_k))\right] \\ &= -\frac{1}{2}t + \frac{1}{2}\beta^2(t),\end{aligned} \tag{3.13}$$

where we have used the result that the limit of the first term is $\lim_{n\to\infty} \sum_k (\beta(t_{k+1}) - \beta(t_k))^2 = t$. The Itô differential of $\beta^2(t)/2$ is analogously

$$\mathrm{d}[\frac{1}{2}\beta^2(t)] = \beta(t)\,\mathrm{d}\beta(t) + \frac{1}{2}\,\mathrm{d}t, \tag{3.14}$$

not $\beta(t)\,\mathrm{d}\beta(t)$ as we might expect. This is a consequence and also a drawback of the selection of the fixed $t_k^* = t_k$.

The general rule for calculating the Itô differentials and thus Itô integrals can be summarized as the following Itô formula, which corresponds to chain rule in ordinary calculus:

Theorem 3.1 (Itô formula)**.** *Assume that $\mathbf{x}(t)$ is an Itô process, and consider an arbitrary (scalar) function $\phi(\mathbf{x}(t), t)$ of the process. Then the Itô differential of ϕ, that is, the Itô SDE for ϕ is given as*

$$\begin{aligned}\mathrm{d}\phi &= \frac{\partial\phi}{\partial t}\,\mathrm{d}t + \sum_i \frac{\partial\phi}{\partial x_i}\,\mathrm{d}x_i + \frac{1}{2}\sum_{ij}\left(\frac{\partial^2\phi}{\partial x_i\,\partial x_j}\right)\mathrm{d}x_i\,\mathrm{d}x_j \\ &= \frac{\partial\phi}{\partial t}\,\mathrm{d}t + (\nabla\phi)^\mathsf{T}\,\mathrm{d}\mathbf{x} + \frac{1}{2}\,\mathrm{tr}\left\{\left(\nabla\nabla^\mathsf{T}\phi\right)\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{x}^\mathsf{T}\right\},\end{aligned} \tag{3.15}$$

provided that the required partial derivatives exist, where the mixed differentials are combined according to the rules

$$\begin{aligned} \mathrm{d}\boldsymbol{\beta}\,\mathrm{d}t &= 0 \\ \mathrm{d}t\,\mathrm{d}\boldsymbol{\beta} &= 0 \\ \mathrm{d}\boldsymbol{\beta}\,\mathrm{d}\boldsymbol{\beta}^{\mathsf{T}} &= \mathbf{Q}\,\mathrm{d}t. \end{aligned} \tag{3.16}$$

Proof. See, for example, Øksendal (2003); Karatzas and Shreve (1991). □

Although the Itô formula above is defined only for scalar ϕ, it obviously works for each of the components of a vector valued function separately and thus also includes the vector case. Also note that every Itô process has a representation as the solution of a SDE of the form

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\,\mathrm{d}t + \mathbf{L}(\mathbf{x}, t)\,\mathrm{d}\boldsymbol{\beta}, \tag{3.17}$$

and an explicit expression for the differential in terms of the functions $\mathbf{f}(\mathbf{x}, t)$ and $\mathbf{L}(\mathbf{x}, t)$ could be derived by substituting the above equation for $\mathrm{d}\mathbf{x}$ in the Itô formula.

The Itô formula can be conceptually derived by a Taylor series expansion:

$$\begin{aligned} \phi(\mathbf{x} + \mathrm{d}\mathbf{x}, t + \mathrm{d}t) = \phi(\mathbf{x}, t) &+ \frac{\partial \phi(\mathbf{x}, t)}{\partial t}\,\mathrm{d}t + \sum_i \frac{\partial \phi(\mathbf{x}, t)}{\partial x_i}\,\mathrm{d}x_i \\ &+ \frac{1}{2}\sum_{ij}\left(\frac{\partial^2 \phi}{\partial x_i\,\partial x_j}\right)\mathrm{d}x_j\,\mathrm{d}x_j + \ldots \end{aligned} \tag{3.18}$$

that is, for the first order in $\mathrm{d}t$ and second order in $\mathrm{d}\mathbf{x}$ we have

$$\begin{aligned} \mathrm{d}\phi &= \phi(\mathbf{x} + \mathrm{d}\mathbf{x}, t + \mathrm{d}t) - \phi(\mathbf{x}, t) \\ &\approx \frac{\partial \phi(\mathbf{x}, t)}{\partial t}\,\mathrm{d}t + \sum_i \frac{\partial \phi(x, t)}{\partial x_i}\,\mathrm{d}x_i + \frac{1}{2}\sum_{ij}\left(\frac{\partial^2 \phi}{\partial x_i\,\partial x_j}\right)\mathrm{d}x_i\,\mathrm{d}x_j. \end{aligned} \tag{3.19}$$

In deterministic case we could ignore the second order and higher order terms, because $\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{x}^{\mathsf{T}}$ would already be of the order $\mathrm{d}t^2$. Thus the deterministic counterpart is

$$\mathrm{d}\phi = \frac{\partial \phi}{\partial t}\,\mathrm{d}t + \frac{\partial \phi}{\partial x}\,\mathrm{d}x. \tag{3.20}$$

But in the stochastic case we know that $\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{x}^{\mathsf{T}}$ is potentially of the order $\mathrm{d}t$, because $\mathrm{d}\boldsymbol{\beta}\,\mathrm{d}\boldsymbol{\beta}^{\mathsf{T}}$ is of the same order. Thus we need to retain the second order term also.

Example 3.1 (Itô differential of $\beta^2(t)/2$)**.** *If we apply the Itô formula to* $\phi(x) = \frac{1}{2}x^2(t)$*, with* $x(t) = \beta(t)$*, where* $\beta(t)$ *is a standard Brownian motion, we get*

$$\begin{aligned} \mathrm{d}\phi &= \beta\,\mathrm{d}\beta + \frac{1}{2}\,\mathrm{d}\beta^2 \\ &= \beta\,\mathrm{d}\beta + \frac{1}{2}\,\mathrm{d}t, \end{aligned} \tag{3.21}$$

as expected.

Example 3.2 (Itô differential of $\sin(\omega\, x)$)**.** *Assume that* $x(t)$ *is the solution to the scalar SDE:*

$$\mathrm{d}x = f(x)\,\mathrm{d}t + \mathrm{d}\beta, \tag{3.22}$$

where $\beta(t)$ *is a Brownian motion with diffusion constant* q *and* $\omega > 0$*. The Itô differential of* $\sin(\omega\, x(t))$ *is then*

$$\begin{aligned} \mathrm{d}[\sin(x)] &= \omega\, \cos(\omega\, x)\,\mathrm{d}x - \frac{1}{2}\omega^2\, \sin(\omega\, x)\,\mathrm{d}x^2 \\ &= \omega\, \cos(\omega\, x)\,[f(x)\,\mathrm{d}t + \mathrm{d}\beta] - \frac{1}{2}\omega^2\, \sin(\omega\, x)\,[f(x)\,\mathrm{d}t + \mathrm{d}\beta]^2 \\ &= \omega\, \cos(\omega\, x)\,[f(x)\,\mathrm{d}t + \mathrm{d}\beta] - \frac{1}{2}\omega^2\, \sin(\omega\, x)\, q\,\mathrm{d}t. \end{aligned} \tag{3.23}$$

3.3 Explicit solutions to linear SDEs

In this section we derive the full solution to a general time-varying linear stochastic differential equation. The time-varying multidimensional SDE is assumed to have the form

$$\mathrm{d}\mathbf{x} = \mathbf{F}(t)\,\mathbf{x}\,\mathrm{d}t + \mathbf{u}(t)\,\mathrm{d}t + \mathbf{L}(t)\,\mathrm{d}\boldsymbol{\beta} \tag{3.24}$$

where $\mathbf{x} \in \mathbb{R}^n$ is the state and $\boldsymbol{\beta} \in \mathbb{R}^s$ is a Brownian motion.

We can now proceed by defining a transition matrix $\boldsymbol{\Psi}(\tau, t)$ in the same way as we did in Equation (1.34). Multiplying the above SDE with the integrating factor $\boldsymbol{\Psi}(t_0, t)$ and rearranging gives

$$\boldsymbol{\Psi}(t_0, t)\,\mathrm{d}\mathbf{x} - \boldsymbol{\Psi}(t_0, t)\,\mathbf{F}(t)\,\mathbf{x}\,\mathrm{d}t = \boldsymbol{\Psi}(t_0, t)\,\mathbf{u}(t)\,\mathrm{d}t + \boldsymbol{\Psi}(t_0, t)\,\mathbf{L}(t)\,\mathrm{d}\boldsymbol{\beta}. \tag{3.25}$$

Applying the Itô formula gives

$$\mathrm{d}[\boldsymbol{\Psi}(t_0, t)\,\mathbf{x}] = -\boldsymbol{\Psi}(t_0, t)\,\mathbf{F}(t)\,\mathbf{x}\,\mathrm{d}t + \boldsymbol{\Psi}(t_0, t)\,\mathrm{d}\mathbf{x}. \tag{3.26}$$

Thus the SDE can be rewritten as

$$\mathrm{d}[\boldsymbol{\Psi}(t_0, t)\,\mathbf{x}] = \boldsymbol{\Psi}(t_0, t)\,\mathbf{u}(t)\,\mathrm{d}t + \boldsymbol{\Psi}(t_0, t)\,\mathbf{L}(t)\,\mathrm{d}\boldsymbol{\beta}, \tag{3.27}$$

where the differential is an Itô differential. Integration (in Itô sense) from t_0 to t gives

$$\boldsymbol{\Psi}(t_0, t)\,\mathbf{x}(t) - \boldsymbol{\Psi}(t_0, t_0)\,\mathbf{x}(t_0) = \int_{t_0}^{t} \boldsymbol{\Psi}(t_0, \tau)\,\mathbf{u}(\tau)\,\mathrm{d}\tau + \int_{t_0}^{t} \boldsymbol{\Psi}(t_0, \tau)\,\mathbf{L}(\tau)\,\mathrm{d}\boldsymbol{\beta}(\tau), \tag{3.28}$$

which can be further written in form

$$\mathbf{x}(t) = \boldsymbol{\Psi}(t, t_0)\,\mathbf{x}(t_0) + \int_{t_0}^{t} \boldsymbol{\Psi}(t, \tau)\,\mathbf{u}(\tau)\,\mathrm{d}\tau + \int_{t_0}^{t} \boldsymbol{\Psi}(t, \tau)\,\mathbf{L}(\tau)\,\mathrm{d}\boldsymbol{\beta}(\tau), \tag{3.29}$$

which is thus the desired full solution.

In the case of a linear time-invariant SDE

$$\mathrm{d}\mathbf{x} = \mathbf{F}\,\mathbf{x}\,\mathrm{d}t + \mathbf{L}\,\mathrm{d}\boldsymbol{\beta}, \tag{3.30}$$

where $\mathbf{F}$ and $\mathbf{L}$ are constant, and $\boldsymbol{\beta}$ has a constant diffusion matrix $\mathbf{Q}$, the solution simplifies to

$$\mathbf{x}(t) = \exp\left(\mathbf{F}\,(t - t_0)\right)\,\mathbf{x}(t_0) + \int_{t_0}^{t} \exp\left(\mathbf{F}\,(t - \tau)\right)\,\mathbf{L}\,\mathrm{d}\boldsymbol{\beta}(\tau). \tag{3.31}$$

By comparing this to Equation (2.35) in Section 2.3, this solution is exactly what we would have expected—it is what we would obtain if we formally replaced $\mathbf{w}(\tau)\,\mathrm{d}\tau$ with $\mathrm{d}\boldsymbol{\beta}(\tau)$ in the deterministic solution. However, it is just because the usage of Itô formula in Equation (3.26) above happened to result in the same result as deterministic differentiation would. In the non-linear case we cannot expect to get the right result with this kind of formal replacement.

Example 3.3 (Solution of the Ornstein–Uhlenbeck process)**.** *The complete solution to the scalar SDE*

$$\mathrm{d}x = -\lambda\,x\,\mathrm{d}t + \mathrm{d}\beta, \qquad x(0) = x_0, \tag{3.32}$$

where $\lambda > 0$ is a given constant and $\beta(t)$ is a Brownian motion is

$$x(t) = \exp(-\lambda\,t)\,x_0 + \int_0^t \exp(-\lambda\,(t - \tau))\,\mathrm{d}\beta(\tau). \tag{3.33}$$

The solution, called the Ornstein–Uhlenbeck process, is illustrated in Figure 3.2.

3.4 Existence and uniqueness of solutions

A solution to a stochastic differential equation is called *strong* if for given Brownian motion $\boldsymbol{\beta}(t)$, it is possible to construct a solution $\mathbf{x}(t)$, which is unique for that given Brownian motion. It means that the whole path of the process is unique

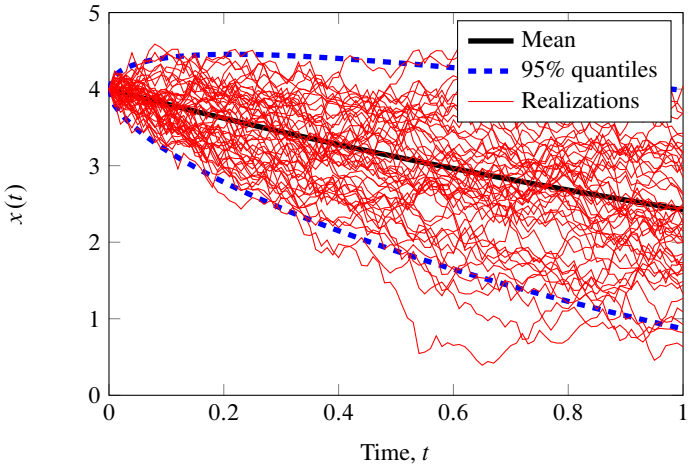


Figure 3.2: Realizations, mean, and 95% quantiles of an Ornstein–Uhlenbeck process.

for a given Brownian motion. Hence strong uniqueness is also called path-wise uniqueness.

The strong uniqueness of a solution to SDE of the general form

$$\mathrm{d}\mathbf{x} = \mathbf{f}(x,t)\,\mathrm{d}t + \mathbf{L}(\mathbf{x},t)\,\mathrm{d}\boldsymbol{\beta}, \quad \mathbf{x}(t_0) = \mathbf{x}_0, \tag{3.34}$$

can be determined using *stochastic Picard's iteration* which is a direct extension of the deterministic Picard's iteration. Thus we first rewrite the equation in integral form

$$\mathbf{x}(t) = \mathbf{x}_0 + \int_{t_0}^{t} \mathbf{f}(x(\tau),\tau)\,\mathrm{d}\tau + \int_{t_0}^{t} \mathbf{L}(x(\tau),\tau)\,\mathrm{d}\boldsymbol{\beta}(\tau). \tag{3.35}$$

Then the solution can be approximated with the following iteration.

Algorithm 3.1 (Stochastic Picard's iteration)**.** *Start from the initial guess* $\boldsymbol{\varphi}_0(t) = \mathbf{x}_0$. *With the given* $\boldsymbol{\beta}$*, compute approximations* $\boldsymbol{\varphi}_1(t), \boldsymbol{\varphi}_2(t), \boldsymbol{\varphi}_3(t), \ldots$ *via the following recursion:*

$$\boldsymbol{\varphi}_{n+1}(t) = \mathbf{x}_0 + \int_{t_0}^{t} \mathbf{f}(\boldsymbol{\varphi}_n(\tau),\tau)\,\mathrm{d}\tau + \int_{t_0}^{t} \mathbf{L}(\boldsymbol{\varphi}_n(\tau),\tau)\,\mathrm{d}\boldsymbol{\beta}(\tau). \tag{3.36}$$

It can be shown that this iteration converges to the exact solution in mean squared sense if both of the functions $\mathbf{f}$ and $\mathbf{L}$ grow at most linearly in $\mathbf{x}$, and they are Lipschitz continuous in the same variable (see, *e.g.*, Øksendal, 2003). If these conditions are met, then there exists a unique strong solution to the SDE.

A solution is called *weak* if it is possible to construct some Brownian motion $\tilde{\boldsymbol{\beta}}(t)$ and a stochastic process $\tilde{\mathbf{x}}(t)$ such that the pair is a solution to the stochastic differential equation. Weak uniqueness means that the probability law of the solution is unique, that is, there cannot be two solutions with different finite-dimensional distributions. An existence of strong solution always implies the existence of a weak solution (every strong solution is also a weak solution), but the

converse is not true. Determination if an equation has a unique weak solution when it does not have a unique strong solution is considerably harder than the criterion for the strong solution.

3.5 Stratonovich calculus

It is also possible to define a stochastic integral in such a way that the chain rule from ordinary calculus is valid. The symmetrized stochastic integral or the *Stratonovich integral* (Stratonovich, 1968) can be defined as follows:

$$\int_{t_0}^{t} \mathbf{L}(\mathbf{x}(t), t) \circ \mathrm{d}\boldsymbol{\beta}(t) = \lim_{n\to\infty} \sum_k \mathbf{L}(\mathbf{x}(t_k^*), t_k^*)\, [\boldsymbol{\beta}(t_{k+1}) - \boldsymbol{\beta}(t_k)], \tag{3.37}$$

where $t_k^* = (t_k + t_{k+1})/2$. The difference is that we do not select the starting point of the interval as the evaluation point, but the middle point. This ensures that the calculation rules of ordinary calculus apply. The disadvantage of the Stratonovich integral over the Itô integral is that the Stratonovich integral is not a martingale, which makes its theoretical analysis harder.

The *Stratonovich stochastic differential equations* (Stratonovich, 1968; Øksendal, 2003) are similar to Itô differential equations, but instead of Itô integrals they involve stochastic integrals in the Stratonovich sense. To distinguish between Itô and Stratonovich stochastic differential equations, the Stratonovich integral is denoted by a small circle before the Brownian differential as follows:

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\, \mathrm{d}t + \mathbf{L}(\mathbf{x}, t) \circ \mathrm{d}\boldsymbol{\beta}. \tag{3.38}$$

The white noise interpretation of SDEs naturally leads to stochastic differential equations in Stratonovich sense. This is because, broadly speaking, discrete-time and smooth approximations of white noise driven differential equations converge to stochastic differential equations in Stratonovich sense, not in Itô sense. However, this result of Wong and Zakai (1965) is strictly true only for scalar SDEs and thus this result should not be extrapolated too far.

A Stratonovich stochastic differential equation can always be converted into an equivalent Itô equation by using simple transformation formulas (Stratonovich, 1968; Øksendal, 2003). If the dispersion term is independent of the state $\mathbf{L}(\mathbf{x}, t) = \mathbf{L}(t)$, then the Itô and Stratonovich interpretations of the stochastic differential equation are the same.

Algorithm 3.2 (Conversion of Stratonovich SDE into Itô SDE)**.** *The following SDE in Stratonovich sense*

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\, \mathrm{d}t + \mathbf{L}(\mathbf{x}, t) \circ \mathrm{d}\boldsymbol{\beta}, \tag{3.39}$$

is equivalent to the following SDE in Itô sense

$$\mathrm{d}\mathbf{x} = \tilde{\mathbf{f}}(\mathbf{x}, t)\, \mathrm{d}t + \mathbf{L}(\mathbf{x}, t)\, \mathrm{d}\boldsymbol{\beta}, \tag{3.40}$$

where

$$\tilde{f}_i(\mathbf{x}, t) = f_i(\mathbf{x}, t) + \frac{1}{2} \sum_{j,k} \frac{\partial L_{ij}(\mathbf{x})}{\partial x_k} L_{kj}(\mathbf{x}). \tag{3.41}$$

Chapter 4

Probability distributions and statistics of SDEs

4.1 Fokker–Planck–Kolmogorov equation

In this section we derive the equation for the probability density of an Itô process $\mathbf{x}(t)$, when the process is defined as the solution to the SDE

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\ \mathrm{d}t + \mathbf{L}(\mathbf{x}, t)\ \mathrm{d}\boldsymbol{\beta}. \tag{4.1}$$

The corresponding probability density is usually denoted as $p(\mathbf{x}(t))$, but in this section, to emphasize that the density is actually function of both $\mathbf{x}$ and t, we will occasionally write it as $p(\mathbf{x}, t)$.

Theorem 4.1 (Fokker–Planck–Kolmogorov equation)**.** *The probability density $p(\mathbf{x}, t)$ of the solution of the SDE in Equation* (4.1) *solves the partial differential equation*

$$\begin{aligned}\frac{\partial p(\mathbf{x}, t)}{\partial t} &= -\sum_i \frac{\partial}{\partial x_i}[f_i(x, t)\ p(\mathbf{x}, t)] \\ &\quad + \frac{1}{2}\sum_{ij} \frac{\partial^2}{\partial x_i\, \partial x_j} \left\{[\mathbf{L}(\mathbf{x}, t)\, \mathbf{Q}\, \mathbf{L}^\mathsf{T}(\mathbf{x}, t)]_{ij}\ p(\mathbf{x}, t)\right\}.\end{aligned} \tag{4.2}$$

This partial differential equation is here called the Fokker–Planck–Kolmogorov (FPK) equation. In physics literature it is often called the Fokker–Planck equation and in stochastics it is the forward Kolmogorov equation, hence the name.

Proof. Let $\phi(\mathbf{x})$ be an arbitrary twice differentiable function. The Itô differential

of $\phi(\mathbf{x}(t))$ is, by the Itô formula, given as follows:

$$\begin{aligned} \mathrm{d}\phi &= \sum_i \frac{\partial \phi}{\partial x_i}\, \mathrm{d}x_i + \frac{1}{2}\sum_{ij}\left(\frac{\partial^2 \phi}{\partial x_i \partial x_j}\right) \mathrm{d}x_i\, \mathrm{d}x_j \\ &= \sum_i \frac{\partial \phi}{\partial x_i}\, f_i(\mathbf{x},t)\, \mathrm{d}t + \sum_i \frac{\partial \phi}{\partial x_i}\, [\mathbf{L}(\mathbf{x},t)\, \mathrm{d}\boldsymbol{\beta}]_i \\ &+ \frac{1}{2}\sum_{ij}\left(\frac{\partial^2 \phi}{\partial x_i \partial x_j}\right) [\mathbf{L}(\mathbf{x},t)\, \mathbf{Q}\, \mathbf{L}^{\mathsf{T}}(\mathbf{x},t)]_{ij}\, \mathrm{d}t. \end{aligned} \tag{4.3}$$

Taking the expectation of both sides with respect to $\mathbf{x}$ and formally dividing by $\mathrm{d}t$ gives the following:

$$\begin{aligned} \frac{\mathrm{d}\,\mathrm{E}[\phi]}{\mathrm{d}t} &= \sum_i \mathrm{E}\left[\frac{\partial \phi}{\partial x_i}\, f_i(\mathbf{x},t)\right] \\ &+ \frac{1}{2}\sum_{ij}\mathrm{E}\left[\left(\frac{\partial^2 \phi}{\partial x_i \partial x_j}\right) [\mathbf{L}(\mathbf{x},t)\, \mathbf{Q}\, \mathbf{L}^{\mathsf{T}}(\mathbf{x},t)]_{ij}\right]. \end{aligned} \tag{4.4}$$

The left hand side can now be written as follows:

$$\begin{aligned} \frac{\mathrm{d}E[\phi]}{\mathrm{d}t} &= \frac{\mathrm{d}}{\mathrm{d}t}\int \phi(\mathbf{x})\, p(\mathbf{x},t)\, \mathrm{d}\mathbf{x} \\ &= \int \phi(\mathbf{x})\, \frac{\partial p(x,t)}{\partial t}\, \mathrm{d}\mathbf{x}. \end{aligned} \tag{4.5}$$

Recall the multidimensional integration by parts formula

$$\int_C \frac{\partial u(\mathbf{x})}{\partial x_i}\, v(\mathbf{x})\, \mathrm{d}\mathbf{x} = \int_{\partial C} u(\mathbf{x})\, v(\mathbf{x})\, n_i\, \mathrm{d}S - \int_C u(\mathbf{x})\, \frac{\partial v(\mathbf{x})}{\partial x_i}\, \mathrm{d}\mathbf{x}, \tag{4.6}$$

where $\mathbf{n}$ is the normal of the boundary ∂C of C and $\mathrm{d}S$ is its area element. If the integration area is whole $\mathbb{R}^n$ and functions $u(\mathbf{x})$ and $v(\mathbf{x})$ vanish at infinity, as is the case here, then the boundary term on the right hand side vanishes and the formula becomes

$$\int \frac{\partial u(\mathbf{x})}{\partial x_i}\, v(\mathbf{x})\, \mathrm{d}\mathbf{x} = -\int u(\mathbf{x})\, \frac{\partial v(\mathbf{x})}{\partial x_i}\, \mathrm{d}\mathbf{x}. \tag{4.7}$$

The term inside the summation of the first term on the right hand side of Equation (4.4) can now be written as

$$\begin{aligned} \mathrm{E}\left[\frac{\partial \phi}{\partial x_i}\, f_i(\mathbf{x},t)\right] &= \int \frac{\partial \phi}{\partial x_i}\, f_i(\mathbf{x},t)\, p(\mathbf{x},t)\, \mathrm{d}\mathbf{x} \\ &= -\int \phi(\mathbf{x})\, \frac{\partial}{\partial x_i}[f_i(\mathbf{x},t)\, p(\mathbf{x},t)]\, \mathrm{d}\mathbf{x}, \end{aligned} \tag{4.8}$$

where we have used the integration by parts formula with $u(\mathbf{x}) = \phi(\mathbf{x})$ and $v(\mathbf{x}) = f_i(\mathbf{x}, t)\ p(\mathbf{x}, t)$. For the term inside the summation sign of the second term we get:

$$\begin{aligned}
&\mathrm{E}\left[\left(\frac{\partial^2 \phi}{\partial x_i \partial x_j}\right)[\mathbf{L}(\mathbf{x}, t)\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}(\mathbf{x}, t)]_{ij}\right] \\
&= \int \left(\frac{\partial^2 \phi}{\partial x_i \partial x_j}\right)[\mathbf{L}(\mathbf{x}, t)\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}(\mathbf{x}, t)]_{ij}\ p(\mathbf{x}, t)\ \mathrm{d}\mathbf{x} \\
&= -\int \left(\frac{\partial \phi}{\partial x_j}\right)\frac{\partial}{\partial x_i}\left\{[\mathbf{L}(\mathbf{x}, t)\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}(\mathbf{x}, t)]_{ij}\ p(\mathbf{x}, t)\right\}\ \mathrm{d}\mathbf{x} \\
&= \int \phi(\mathbf{x})\frac{\partial^2}{\partial x_i\ \partial x_j}\left\{[\mathbf{L}(\mathbf{x}, t)\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}(\mathbf{x}, t)]_{ij}\ p(\mathbf{x}, t)\right\}\ \mathrm{d}\mathbf{x},
\end{aligned} \tag{4.9}$$

where we have first used the integration by parts formula with $u(\mathbf{x}) = \partial\phi(\mathbf{x})/\partial x_j$, $v(\mathbf{x}) = [\mathbf{L}(\mathbf{x}, t)\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}(\mathbf{x}, t)]_{ij}\ p(\mathbf{x}, t)$ and then again with $u(\mathbf{x}) = \phi(\mathbf{x})$, $v(\mathbf{x}) = \frac{\partial}{\partial x_i}\{[\mathbf{L}(\mathbf{x}, t)\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}(\mathbf{x}, t)]_{ij}\ p(\mathbf{x}, t)\}$.

If we substitute Equations (4.5), (4.8), and (4.9) into (4.4), we get:

$$\begin{aligned}
&\int \phi(\mathbf{x})\,\frac{\partial p(\mathbf{x}, t)}{\partial t}\ \mathrm{d}\mathbf{x} \\
&\quad = -\sum_i \int \phi(\mathbf{x})\,\frac{\partial}{\partial x_i}[f_i(\mathbf{x}, t)\ p(\mathbf{x}, t)]\ \mathrm{d}\mathbf{x} \\
&\quad + \frac{1}{2}\sum_{ij} \int \phi(\mathbf{x})\frac{\partial^2}{\partial x_i\ \partial x_j}\{[\mathbf{L}(\mathbf{x}, t)\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}(\mathbf{x}, t)]_{ij}\ p(\mathbf{x}, t)\}\ \mathrm{d}\mathbf{x},
\end{aligned} \tag{4.10}$$

which can also be written as

$$\begin{aligned}
&\int \phi(\mathbf{x})\Big[\frac{\partial p(\mathbf{x}, t)}{\partial t} + \sum_i \frac{\partial}{\partial x_i}[f_i(\mathbf{x}, t)\ p(\mathbf{x}, t)] \\
&- \frac{1}{2}\sum_{ij} \frac{\partial^2}{\partial x_i\ \partial x_j}\{[\mathbf{L}(\mathbf{x}, t)\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}(\mathbf{x}, t)]_{ij}\ p(\mathbf{x}, t)\}\Big]\ \mathrm{d}\mathbf{x} = 0.
\end{aligned} \tag{4.11}$$

The only way that this equation can be true for an arbitrary $\phi(\mathbf{x})$ is if the term in the brackets vanishes, which gives the FPK equation. □

Example 4.1 (Diffusion equation)**.** *In Example 2.1 we derived the diffusion equation by considering random Brownian movement occurring during small time intervals. Note that Brownian motion can be defined as a solution to the SDE*

$$\mathrm{d}x = \mathrm{d}\beta. \tag{4.12}$$

If we set the diffusion constant of the Brownian motion to be $q = 2\,D$, then the FPK reduces to

$$\frac{\partial p}{\partial t} = D\,\frac{\partial^2 p}{\partial x^2}, \tag{4.13}$$

which is the same result as in Equation (2.7)*.*

4.2 Operator formulation of the FPK equation

First note that if we define a linear operator $\mathcal{A}^*$ (operating on some function $\bullet$) as

$$\begin{aligned}\mathcal{A}^*(\bullet) = &-\sum_i \frac{\partial}{\partial x_i}[f_i(x,t)\,(\bullet)] \\ &+ \frac{1}{2}\sum_{ij}\frac{\partial^2}{\partial x_i\,\partial x_j}\{[\mathbf{L}(\mathbf{x},t)\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}(\mathbf{x},t)]_{ij}\,(\bullet)\},\end{aligned} \tag{4.14}$$

then Fokker–Planck–Kolmomogorov equation can be written compactly as

$$\frac{\partial p}{\partial t} = \mathcal{A}^* p. \tag{4.15}$$

which is just linear differential equation—however, an infinite-dimensional one.

This operator formulation allows us to gain insight on what actually happened in the FPK derivation in the previous section. Let us now define the L_2 inner product between two functions ϕ and φ as follows:

$$\langle \phi, \varphi \rangle = \int \phi(\mathbf{x})\,\varphi(\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{4.16}$$

The expectation of a function $\phi(\mathbf{x}(t))$ can now be written in terms of the inner product as follows:

$$\mathrm{E}[\phi(\mathbf{x}(t))] = \langle \phi, p \rangle, \tag{4.17}$$

where $p = p(\mathbf{x}, t)$. This also means that Equation (4.4), which was derived from the Itô formula, can be compactly written as

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle \phi, p \rangle = \langle \mathcal{A}\,\phi, p \rangle, \tag{4.18}$$

where

$$\begin{aligned}\mathcal{A}(\bullet) = &\sum_i \mathrm{E}\left[f_i(\mathbf{x},t)\,\frac{\partial(\bullet)}{\partial x_i}\right] \\ &+ \frac{1}{2}\sum_{ij}\mathrm{E}\left[[\mathbf{L}(\mathbf{x},t)\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}(\mathbf{x},t)]_{ij}\,\left(\frac{\partial^2(\bullet)}{\partial x_i\,\partial x_j}\right)\right].\end{aligned} \tag{4.19}$$

Recall that the adjoint of an operator $\mathcal{A}$—with respect to the given inner product—is defined to be an operator $\mathcal{A}^*$ such that for all ϕ and φ we have $\langle \mathcal{A}\,\phi, \varphi \rangle = \langle \phi, \mathcal{A}^*\,\varphi \rangle$. In terms of the adjoint operator we can now write Equation (4.18) as

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle \phi, p \rangle = \langle \phi, \mathcal{A}^*\,p \rangle. \tag{4.20}$$

As ϕ is independent of time, this can also be written as

$$\langle \phi, \frac{\partial p}{\partial t} \rangle = \langle \phi, \mathcal{A}^*\,p \rangle. \tag{4.21}$$

Because ϕ can be arbitrary, the above can only be true if in fact

$$\frac{\partial p}{\partial t} = \mathcal{A}^*\, p. \tag{4.22}$$

It now turns out that the adjoint of the operator $\mathcal{A}$ in Equation (4.19) is exactly the operator in Equation (4.14) and hence the above equation is in fact the Fokker–Planck–Kolmogorov equation. What we did in the previous section is that we used brute-force integration by parts to derive the adjoint of the operator $\mathcal{A}$. We could also have used the properties of the adjoints directly as is illustrated in the following example.

Example 4.2 (Operator adjoint derivation of FKP)**.** *Let us consider a one-dimensional SDE*

$$\mathrm{d}x = f(x)\,\mathrm{d}t + L(x)\,\mathrm{d}\beta, \tag{4.23}$$

in which case the operator $\mathcal{A}$ takes the form

$$\mathcal{A} = f(x)\,\frac{\partial}{\partial x} + \frac{1}{2}\,L^2(x)\,q\,\frac{\partial^2}{\partial x^2}. \tag{4.24}$$

Now recall the following L_2 adjoint computation rules:

- *The operation of multiplication with a function $f(x)$ is its own adjoint (i.e., the operator is self-adjoint).*
- *The operation of differentiation obeys $\left(\frac{\partial}{\partial x}\right)^* = -\frac{\partial}{\partial x}$ and hence the second derivative operator is self-adjoint $\left(\frac{\partial^2}{\partial x^2}\right)^* = \frac{\partial^2}{\partial x^2}$.*
- *The adjoint of a sum is $(\mathcal{A}_1 + \mathcal{A}_2)^* = \mathcal{A}_1^* + \mathcal{A}_2^*$ and the product of two operators is $(\mathcal{A}_1\,\mathcal{A}_2)^* = \mathcal{A}_2^*\,\mathcal{A}_1^*$.*

Thus we get

$$\begin{aligned}
\langle \mathcal{A}\,\phi, p\rangle &= \langle \left[f(x)\,\frac{\partial}{\partial x} + \frac{1}{2}\,L^2(x)\,q\,\frac{\partial^2}{\partial x^2}\right]\phi, p\rangle \\
&= \langle f(x)\,\frac{\partial}{\partial x}\,\phi, p\rangle + \frac{1}{2}\,q\,\langle L^2(x)\,\frac{\partial^2}{\partial x^2}\,\phi, p\rangle \\
&= \langle \frac{\partial}{\partial x}\,\phi, f(x)\,p\rangle + \frac{1}{2}\,q\,\langle \frac{\partial^2}{\partial x^2}\,\phi, L^2(x)\,p\rangle \\
&= \langle \phi, \left[-\frac{\partial}{\partial x}\right] f(x)\,p\rangle + \frac{1}{2}\,q\,\langle \phi, \frac{\partial^2}{\partial x^2}\,L^2(x)\,p\rangle \\
&= \langle \phi, -\frac{\partial}{\partial x}\,f(x)\,p\rangle + \frac{1}{2}\,q\,\langle \phi, \frac{\partial^2}{\partial x^2}\,L^2(x)\,p\rangle \\
&= \langle \phi, -\frac{\partial}{\partial x}\,f(x)\,p + \frac{1}{2}\,q\,\frac{\partial^2}{\partial x^2}\,L^2(x)\,p\rangle,
\end{aligned} \tag{4.25}$$

where we can thus recover the adjoint operator

$$\mathcal{A}^*(\bullet) = -\frac{\partial}{\partial x}\,\left[f(x)\,(\bullet)\right] + \frac{1}{2}\,q\,\frac{\partial^2}{\partial x^2}\,\left[L^2(x)\,(\bullet)\right]. \tag{4.26}$$

4.3 Markov properties and transition densities of SDEs

In this section the aim is to discuss the Markovian property of Itô processes and the corresponding transition kernels. We denote the *history* of the Itô process $\mathbf{x}(t)$ up to the time t as

$$\mathcal{X}_t = \{\mathbf{x}(\tau) \,:\, 0 \leq \tau \leq t\}. \tag{4.27}$$

More formally, the history of an Itô process should not be defined through its explicit path, but via the *sigma-algebra* generated by it (see, *e.g.*, Øksendal, 2003). The history as function of increasing t is then an object called *filtration*, which means an increasing family of sigma-algebras. However, for pedagogical reasons we simply talk about the history of an Itô process.

The definition of a *Markov process* is the following:

Definition 4.1 (Markov process)**.** *A stochastic process* $\mathbf{x}(t)$ *is a* Markov process *if its future is independent of its past given the present:*

$$p(\mathbf{x}(t) \mid \mathcal{X}_s) = p(\mathbf{x}(t) \mid \mathbf{x}(s)), \quad \text{for all } t \geq s. \tag{4.28}$$

It turns out that all Itô processes, that is, solutions to Itô stochastic differential equations are Markov processes. The proof of this can be found, for example, in Øksendal (2003, Theorem 7.1.2). This means that the all Itô processes are, in probabilistic sense, completely characterized by the two-parameter family of transition densities $p(\mathbf{x}(t) \mid \mathbf{x}(s))$. The transition density is also a solution to the Fokker–Planck–Kolmogorov equation with a degenerate initial condition concentrated on $\mathbf{x}(s)$ at time s.

Theorem 4.2 (Transition density of an SDE)**.** *The transition density* $p(\mathbf{x}(t) \mid \mathbf{x}(s))$ *of the SDE* (4.1)*, where* $t \geq s$*, is the solution to the Fokker–Planck–Kolmogorov equation* (4.2) *with the initial condition* $p(\mathbf{x}, s) = \delta(\mathbf{x} - \mathbf{x}(s))$.

Once we know the transition densities of an SDE, we can also use Markov properties to form an explicit formula for the *finite-dimensional distributions* of the SDE.

Remark 4.1 (Finite-dimensional distributions of SDEs)**.** *For an arbitrary finite-set of time indices* $t_0 < t_1 < \cdots < t_n$ *the joint distribution of the values of the process (i.e., the finite-dimensional distribution) is*

$$p(\mathbf{x}(t_0), \mathbf{x}(t_1), \ldots, \mathbf{x}(t_n)) = p(\mathbf{x}(t_0)) \prod_{i=1}^{n} p(\mathbf{x}(t_i) \mid \mathbf{x}(t_{i-1})). \tag{4.29}$$

The above result is extremely important in Bayesian filtering theory (Särkkä, 2013), because it states that a Bayesian filtering problem on an SDE model with discrete time measurements can always be converted into an equivalent discrete-time Bayesian filtering problem. We will return to this in the later chapters.

Remark 4.2 (Chapman–Kolmogorov equation)**.** *The Markov property also implies that the transition densities have the following group property which says that for any three time instants $t_1 < t_2 < t_3$ we have the Chapman–Kolmogorov equation*

$$p(\mathbf{x}(t_3) \mid \mathbf{x}(t_1)) = \int p(\mathbf{x}(t_3) \mid \mathbf{x}(t_2))\, p(\mathbf{x}(t_2) \mid \mathbf{x}(t_1))\, \mathrm{d}\mathbf{x}(t_2). \tag{4.30}$$

Although in the present formulation the above equation follows from the FPK equation, it is also possible to derive the FPK equation from it as was done, for example, in Jazwinski (1970).

4.4 Mean and covariance of SDEs

In Section 4.1 we derived the Fokker–Planck–Kolmogorov (FPK) equation which, in principle, is the complete probabilistic description of the state. The mean, covariance and other moments of the state distribution can be derived from its solution. However, we are often primarily interested in the mean and covariance of the distribution and would like to avoid solving the FPK equation as an intermediate step.

If we take a look at the Equation (4.4) in Section 4.1, we can see that it can be interpreted as equation for the general moments of the state distribution. This equation can be generalized to time dependent $\phi(\mathbf{x}, t)$ by including the time derivative:

$$\begin{aligned}\frac{\mathrm{d}\,\mathrm{E}[\phi]}{\mathrm{d}t} &= \mathrm{E}\left[\frac{\partial \phi}{\partial t}\right] + \sum_i \mathrm{E}\left[\frac{\partial \phi}{\partial x_i}\, f_i(\mathbf{x}, t)\right] \\ &\quad + \frac{1}{2}\sum_{ij} \mathrm{E}\left[\left(\frac{\partial^2 \phi}{\partial x_i\, \partial x_j}\right) [\mathbf{L}(\mathbf{x}, t)\, \mathbf{Q}\, \mathbf{L}^{\mathsf{T}}(\mathbf{x}, t)]_{ij}\right].\end{aligned} \tag{4.31}$$

If we select the function as $\phi(x, t) = x_u$, then the Equation (4.31) reduces to

$$\frac{\mathrm{d}\,\mathrm{E}[x_u]}{\mathrm{d}t} = \mathrm{E}\left[f_u(\mathbf{x}, t)\right] \tag{4.32}$$

which can be seen as the differential equation for the components of the mean of the state. If we denote the mean function as $\mathbf{m}(t) = \mathrm{E}[\mathbf{x}(t)]$ and select the function as $\phi(\mathbf{x}, t) = x_u\, x_v - m_u(t)\, m_v(t)$, then Equation (4.31) gives

$$\begin{aligned}&\frac{\mathrm{d}\,\mathrm{E}[x_u\, x_v - m_u(t)\, m_v(t)]}{\mathrm{d}t} \\ &= \mathrm{E}\left[(x_v - m_v(t))\, f_u(x, t)\right] + \mathrm{E}\left[(x_u - m_u(v))\, f_v(x, t)\right] \\ &\quad + [\mathbf{L}(\mathbf{x}, t)\, \mathbf{Q}\, \mathbf{L}^{\mathsf{T}}(\mathbf{x}, t)]_{uv}.\end{aligned} \tag{4.33}$$

If we denote the covariance as $\mathbf{P}(t) = \mathrm{E}[(\mathbf{x}(t) - \mathbf{m}(t))\, (\mathbf{x}(t) - \mathbf{m}(t))^{\mathsf{T}}]$, then

Equations (4.32) and (4.33) can be written in the following matrix form:

$$\frac{\mathrm{d}\mathbf{m}}{\mathrm{d}t} = \mathrm{E}\left[\mathbf{f}\left(\mathbf{x}, t\right)\right] \tag{4.34}$$

$$\begin{aligned}\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}t} &= \mathrm{E}\left[\mathbf{f}\left(\mathbf{x}, t\right)\left(\mathbf{x} - \mathbf{m}\right)^{\mathsf{T}}\right] + \mathrm{E}\left[\left(\mathbf{x} - \mathbf{m}\right)\mathbf{f}^{\mathsf{T}}(\mathbf{x}, t)\right] \\ &\quad + \mathrm{E}\left[\mathbf{L}(\mathbf{x}, t)\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}(\mathbf{x}, t)\right],\end{aligned} \tag{4.35}$$

which are the differential equations for the mean and covariance of the state. However, these equations cannot be used in practice as such, because the expectations should be taken with respect to the actual distribution of the state—which is the solution to the FPK equation. Only in the Gaussian case the first two moments actually characterize the solution. Even though we cannot use these equations as such in the non-linear case, they provide a useful starting point for forming Gaussian approximations to SDEs.

Example 4.3 (Moments of an Ornstein–Uhlenbeck process)**.** *Let us again consider the Ornstein–Uhlenbeck process which we solved in Example 3.3:*

$$\mathrm{d}x = -\lambda\, x\, \mathrm{d}t + \mathrm{d}\beta, \qquad x(0) = x_0, \tag{4.36}$$

where $\lambda > 0$ and $\beta(t)$ is a Brownian motion with diffusion constant q. We have $f(x) = -\lambda\, x$ and thus

$$\begin{aligned}\mathrm{E}[f(x)] &= -\lambda\ \mathrm{E}[x] = -\lambda\, m \\ \mathrm{E}[f(x)\,(x - m)] &= \mathrm{E}[-\lambda\, x\,(x - m)] = -\lambda\ \mathrm{E}[(x - m)^2] = -\lambda\, P.\end{aligned} \tag{4.37}$$

The differential equations for the mean and variance are thus given as

$$\begin{aligned}\frac{\mathrm{d}m}{\mathrm{d}t} &= -\lambda\, m, \\ \frac{\mathrm{d}P}{\mathrm{d}t} &= -2\lambda\, P + q,\end{aligned} \tag{4.38}$$

with the initial conditions $m(0) = x_0$, $P(0) = 0$. Because the solution of the Ornstein–Uhlenbeck process is a Gaussian process, these first two moments characterize the whole state distribution which is

$$p(x, t) \triangleq p(x(t)) = \mathrm{N}(x(t) \mid m(t), P(t)). \tag{4.39}$$

4.5 Higher order moments of SDEs

It is also possible to derive differential equations for the higher order moments of SDEs. However, the required number of equations quickly becomes huge, because if the state dimension is n, the number of independent third moments is cubic, n^3, in the number of state dimension. The number of fourth order moments is quartic,

n^4, and so on. The general moments equations can be found, for example, in the book of Socha (2008).

To illustrate the idea, let's consider the scalar SDE

$$\mathrm{d}x = f(x)\ \mathrm{d}t + L(x)\ \mathrm{d}\beta. \tag{4.40}$$

Recall that the expectation of an arbitrary twice differentiable function $\phi(x)$ satisfies

$$\frac{\mathrm{d}\,\mathrm{E}[\phi(x)]}{\mathrm{d}t} = \mathrm{E}\left[\frac{\partial \phi(x)}{\partial x}\, f(x)\right] + \frac{q}{2}\,\mathrm{E}\left[\frac{\partial^2 \phi(x)}{\partial x^2}\, L^2(x)\right]. \tag{4.41}$$

If we apply this to $\phi(x) = x^n$, where $n \geq 2$, we get

$$\frac{\mathrm{d}\,\mathrm{E}[x^n]}{\mathrm{d}t} = n\ \mathrm{E}[x^{n-1}\ f(x,t)] + \frac{q}{2}\, n\,(n-1)\ \mathrm{E}[x^{n-2}\ L^2(x)], \tag{4.42}$$

which, in principle, gives the equations for the third order moments, fourth order moments and so on. It is also possible to derive similar differential equations for the central moments, cumulants, or quasi-moments.

However, unless $f(x)$ and $L(x)$ are linear (or affine) functions, the equation for the nth order moment depends on the moments of higher order (greater than n). Thus in order to actually compute the required expectations, we would need to integrate an infinite number of moment equations, which is impossible in practice. This problem can be solved by using moment closure methods which typically are based on replacing the higher order moments (or cumulants or quasi-moments) with suitable approximations. For example, it is possible to set the cumulants above a certain order to zero, or to approximate the moments/cumulants/quasi-moments with their steady state values (Socha, 2008).

In the scalar case, given a set of moments, cumulants or quasi-moments, it is possible to form a distribution which has these moments/cumulants/quasi-moments, for example, as the maximum entropy distribution. Unfortunately, in the multidimensional case the situation is much more complicated.

4.6 Mean, covariance, transition density of linear SDEs

Let's now consider a linear stochastic differential equation of the general form

$$\mathrm{d}\mathbf{x} = \mathbf{F}(t)\,\mathbf{x}(t)\,\mathrm{d}t + \mathbf{u}(t)\ \mathrm{d}t + \mathbf{L}(t)\ \mathrm{d}\boldsymbol{\beta}(t), \tag{4.43}$$

where the initial conditions are $\mathbf{x}(t_0) \sim \mathrm{N}(\mathbf{m}_0, \mathbf{P}_0)$, $\mathbf{F}(t)$ and $\mathbf{L}(t)$ are matrix valued functions of time, $\mathbf{u}(t)$ is a vector valued function of time and $\boldsymbol{\beta}(t)$ is a Brownian motion with diffusion matrix $\mathbf{Q}$.

The mean and covariance can be solved from the Equations (4.34) and (4.35), which in this case reduce to

$$\begin{aligned} \frac{\mathrm{d}\mathbf{m}(t)}{\mathrm{d}t} &= \mathbf{F}(t)\,\mathbf{m}(t) + \mathbf{u}(t) \\ \frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t} &= \mathbf{F}(t)\,\mathbf{P}(t) + \mathbf{P}(t)\,\mathbf{F}^{\mathsf{T}}(t) + \mathbf{L}(t)\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}(t) \end{aligned} \tag{4.44}$$

with the initial conditions $\mathbf{m}_0(t_0) = \mathbf{m}_0$ and $\mathbf{P}(t_0) = \mathbf{P}_0$. The general solutions to these differential equations are

$$\mathbf{m}(t) = \boldsymbol{\Psi}(t, t_0)\,\mathbf{m}(t_0) + \int_{t_0}^{t} \boldsymbol{\Psi}(t, \tau)\,\mathbf{u}(\tau)\,\mathrm{d}\tau \tag{4.45}$$

$$\mathbf{P}(t) = \boldsymbol{\Psi}(t, t_0)\,\mathbf{P}(t_0)\,\boldsymbol{\Psi}^{\mathsf{T}}(t, t_0) + \int_{t_0}^{t} \boldsymbol{\Psi}(t, \tau)\,\mathbf{L}(\tau)\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}(\tau)\,\boldsymbol{\Psi}^{\mathsf{T}}(t, \tau)\,\mathrm{d}\tau, \tag{4.46}$$

which could also be obtained by computing the mean and covariance of the explicit solution in Equation (3.29).

Because the solution is a linear transformation of the Brownian motion, which is a Gaussian process, the solution is Gaussian

$$p(\mathbf{x}, t) \triangleq p(\mathbf{x}(t)) = \mathrm{N}(\mathbf{x}(t) \,|\, \mathbf{m}(t), \mathbf{P}(t)), \tag{4.47}$$

which can be verified by checking that this distribution indeed solves the corresponding Fokker–Planck–Kolmogorov equation (4.2). Furthermore, the transition density can be recovered by formally using the initial condition $\mathbf{m}(s) = \mathbf{x}(s)$, $\mathbf{P}(s) = 0$, which gives

$$p(\mathbf{x}(t) \mid \mathbf{x}(s)) = \mathrm{N}(\mathbf{x}(t) \,|\, \mathbf{m}(t \mid s), \mathbf{P}(t \mid s)), \tag{4.48}$$

where

$$\begin{aligned} \mathbf{m}(t \mid s) &= \boldsymbol{\Psi}(t, s)\,\mathbf{x}(s) + \int_{s}^{t} \boldsymbol{\Psi}(t, \tau)\,\mathbf{u}(\tau)\,\mathrm{d}\tau \\ \mathbf{P}(t \mid s) &= \int_{s}^{t} \boldsymbol{\Psi}(t, \tau)\,\mathbf{L}(\tau)\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}(\tau)\,\boldsymbol{\Psi}^{\mathsf{T}}(t, \tau)\,\mathrm{d}\tau. \end{aligned} \tag{4.49}$$

It is now useful to note that the above implies that the original linear SDE is (weakly, in distribution) equivalent to the following discrete-time system:

$$\mathbf{x}(t_{k+1}) = \mathbf{A}_k\,\mathbf{x}(t_k) + \mathbf{u}_k + \mathbf{q}_k, \qquad \mathbf{q}_k \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_k), \tag{4.50}$$

where

$$\mathbf{A}_k \triangleq \boldsymbol{\Psi}(t_{k+1}, t_k), \tag{4.51}$$

$$\mathbf{u}_k \triangleq \int_{t_k}^{t_{k+1}} \boldsymbol{\Psi}(t, \tau)\,\mathbf{u}(\tau)\,\mathrm{d}\tau \tag{4.52}$$

$$\boldsymbol{\Sigma}_k \triangleq \boldsymbol{\Sigma}(t_{k+1}, t_k) = \int_{t_k}^{t_{k+1}} \boldsymbol{\Psi}(t, \tau)\,\mathbf{L}(\tau)\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}(\tau)\,\boldsymbol{\Psi}^{\mathsf{T}}(t, \tau)\,\mathrm{d}\tau, \tag{4.53}$$

which is sometimes called the *equivalent discretization* of the SDEs in Kalman filtering context (*cf.* Grewal and Andrews, 2001; Särkkä, 2006, 2013).

4.7 Linear time-invariant SDEs and matrix fractions

In the case of a linear time-invariant SDE

$$\mathrm{d}\mathbf{x} = \mathbf{F}\,\mathbf{x}(t)\,\mathrm{d}t + \mathbf{L}\,\mathrm{d}\boldsymbol{\beta}(t), \tag{4.54}$$

the mean and covariance are also given by Equations (4.45) and (4.46), which now take the form

$$\begin{aligned} \frac{\mathrm{d}\mathbf{m}(t)}{\mathrm{d}t} &= \mathbf{F}\,\mathbf{m}(t) \\ \frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t} &= \mathbf{F}\,\mathbf{P}(t) + \mathbf{P}(t)\,\mathbf{F}^\mathsf{T} + \mathbf{L}\,\mathbf{Q}\,\mathbf{L}^\mathsf{T}. \end{aligned} \tag{4.55}$$

Thus the only differences are that the matrices $\mathbf{F}$ and $\mathbf{L}$ are constant, and there is no input. In this LTI SDE case the transition matrix is the matrix exponential function $\boldsymbol{\Psi}(t,\tau) = \exp(\mathbf{F}\,(t-\tau))$ and the solutions to the differential equations can be obtained by simple substitution of these special cases to Equations (4.45) and (4.46). The transition density can be obtained in an analogous manner.

Fortunately, the corresponding equivalent discrete system now takes a particularly simple form

$$\mathbf{x}(t_{k+1}) = \mathbf{A}_k\,\mathbf{x}(t_k) + \mathbf{q}_k, \quad \mathbf{q}_k \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_k), \tag{4.56}$$

where $\Delta t_k = t_{k+1} - t_k$ and

$$\mathbf{A}_k \triangleq \mathbf{A}(\Delta t_k) = \exp(\mathbf{F}\,\Delta t_k) \tag{4.57}$$

$$\boldsymbol{\Sigma}_k \triangleq \boldsymbol{\Sigma}(\Delta t_k) = \int_0^{\Delta t_k} \exp(\mathbf{F}\,(\Delta t_k - \tau))\,\mathbf{L}\,\mathbf{Q}\,\mathbf{L}^\mathsf{T}\,\exp(\mathbf{F}\,(\Delta t_k - \tau))^\mathsf{T}\,\mathrm{d}\tau. \tag{4.58}$$

These equations can often be found in tracking literature (Bar-Shalom et al., 2001; Grewal and Andrews, 2001; Särkkä, 2006, 2013), because they are extremely useful in converting continuous-discrete Kalman filtering problems into equivalent discrete-time Kalman filtering problems. A typical example of a model in that context is the following.

Example 4.4 (Discretized Wiener velocity model)**.** *The Wiener velocity model (see, e.g., Bar-Shalom et al., 2001; Särkkä, 2006) is a typical model found in target tracking, where the velocity (the first derivative of the process) is modeled as a Wiener process, that is, as a Brownian motion. In white noise interpretation this means that the acceleration (i.e., the second derivative) is a white noise process with spectral density q:*

$$\frac{\mathrm{d}^2 x(t)}{\mathrm{d}t^2} = w(t). \tag{4.59}$$

In more rigorous Itô SDE form this model can be written as

$$\begin{pmatrix} \mathrm{d}x_1 \\ \mathrm{d}x_2 \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}}_{\mathbf{F}} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mathrm{d}t + \underbrace{\begin{pmatrix} 0 \\ 1 \end{pmatrix}}_{\mathbf{L}} \mathrm{d}\beta(t), \tag{4.60}$$

where $\beta(t)$ is a Brownian motion with diffusion coefficient q, $x_1(t) \triangleq x(t)$ is the actual process, and $x_2(t)$ is its derivative.

Now the matrices of the equivalent discrete-time model are given as follows (notice that $\mathbf{F}$ is a nilpotent matrix such that $\mathbf{F}^n = 0$ for $n > 1$):

$$\mathbf{A}(\Delta t) = \exp(\mathbf{F}\,\Delta t) = \mathbf{I} + \mathbf{F}\,\Delta + \underbrace{\frac{1}{2!}\mathbf{F}^2\,\Delta t^2 + \cdots}_{=0} = \begin{pmatrix} 1 & \Delta t \\ 0 & 1 \end{pmatrix}$$

$$\mathbf{\Sigma}(\Delta t) = \int_0^{\Delta t} \begin{pmatrix} 1 & \Delta t - \tau \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & q \end{pmatrix} \begin{pmatrix} 1 & \Delta t - \tau \\ 0 & 1 \end{pmatrix}^{\mathsf{T}} \mathrm{d}\tau = \begin{pmatrix} \frac{1}{3}\Delta t^3 & \frac{1}{2}\Delta t^2 \\ \frac{1}{2}\Delta t^2 & \Delta t \end{pmatrix} q. \tag{4.61}$$

Example 4.5 (Discretization of the car model)**.** *A 2d-dimensional version of the above Wiener velocity model was already presented in Example 2.5 for the purpose of modeling the movement of a car. The same model was also used in a Kalman filtering and smoothing context, for example, in Särkkä (2013). The corresponding discrete-time model matrices now become:*

$$\mathbf{A}(\Delta t) = \begin{pmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{\Sigma}(\Delta t) = \begin{pmatrix} \frac{q_1^c\,\Delta t^3}{3} & 0 & \frac{q_1^c\,\Delta t^2}{2} & 0 \\ 0 & \frac{q_2^c\,\Delta t^3}{3} & 0 & \frac{q_2^c\,\Delta t^2}{2} \\ \frac{q_1^c\,\Delta t^2}{2} & 0 & q_1^c\,\Delta t & 0 \\ 0 & \frac{q_2^c\,\Delta t^2}{2} & 0 & q_2^c\,\Delta t \end{pmatrix}.$$

A convenient numerical method for solving the covariance $\mathbf{P}(t)$ from Equations (4.55) is by using matrix fractions (see, *e.g.*, Stengel, 1994; Grewal and Andrews, 2001; Särkkä, 2006). If we define matrices $\mathbf{C}(t)$ and $\mathbf{D}(t)$ such that $\mathbf{P}(t) = \mathbf{C}(t)\,\mathbf{D}^{-1}(t)$, it is easy to show that $\mathbf{P}$ solves the matrix Lyapunov differential equation

$$\frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t} = \mathbf{F}\,\mathbf{P}(t) + \mathbf{P}(t)\,\mathbf{F}^{\mathsf{T}} + \mathbf{L}\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}, \tag{4.62}$$

if the matrices $\mathbf{C}(t)$ and $\mathbf{D}(t)$ solve the differential equation

$$\begin{pmatrix} \mathrm{d}\mathbf{C}(t)/\,\mathrm{d}t \\ \mathrm{d}\mathbf{D}(t)/\,\mathrm{d}t \end{pmatrix} = \begin{pmatrix} \mathbf{F} & \mathbf{L}\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}} \\ \mathbf{0} & -\mathbf{F}^{\mathsf{T}} \end{pmatrix} \begin{pmatrix} \mathbf{C}(t) \\ \mathbf{D}(t) \end{pmatrix}, \tag{4.63}$$

and $\mathbf{P}(t_0) = \mathbf{C}(t_0)\,\mathbf{D}(t_0)^{-1}$. We can select, for example,

$$\mathbf{C}(t_0) = \mathbf{P}(t_0) \tag{4.64}$$

$$\mathbf{D}(t_0) = \mathbf{I}. \tag{4.65}$$

Because the differential equation (4.63) is linear and time-invariant, it can be solved using the matrix exponential function:

$$\begin{pmatrix} \mathbf{C}(t) \\ \mathbf{D}(t) \end{pmatrix} = \exp\left\{ \begin{pmatrix} \mathbf{F} & \mathbf{L}\mathbf{Q}\mathbf{L}^\top \\ \mathbf{0} & -\mathbf{F}^\top \end{pmatrix} t \right\} \begin{pmatrix} \mathbf{C}(t_0) \\ \mathbf{D}(t_0) \end{pmatrix}. \tag{4.66}$$

The final solution is then given as $\mathbf{P}(t) = \mathbf{C}(t)\,\mathbf{D}^{-1}(t)$. This is useful, because now both the mean and covariance can be solved via simple matrix exponential function computation, which allows for easy numerical treatment.

In practical Kalman filter we are often interested in forming the matrices $\mathbf{A}(\Delta t)$ and $\mathbf{\Sigma}(\Delta t)$ in Equations (4.57) and (4.58) by numerical means. This is because these numerical matrices can then be directly used in a discrete-time Kalman filter to infer the state of the SDE at a discrete set of time instants (see, *e.g.*, Särkkä, 2013). The numerical computation of $\mathbf{A}(\Delta t)$ is easy, because it is just a matrix exponential for which good numerical computation methods are available. However, the integral expression for $\mathbf{\Sigma}(\Delta t)$ is more problematic in numerical point of view.

It turns out that the matrix fractions can also be used reduce the computation of the matrix $\mathbf{\Sigma}(\Delta t)$ into a simple matrix exponential (Särkkä, 2006). The trick is that the matrix is also the solution to the differential equation

$$\frac{\mathrm{d}\mathbf{\Sigma}(t)}{\mathrm{d}t} = \mathbf{F}\,\mathbf{\Sigma}(t) + \mathbf{\Sigma}(t)\,\mathbf{F}^\top + \mathbf{L}\,\mathbf{Q}\,\mathbf{L}^\top, \quad \mathbf{\Sigma}(0) = \mathbf{0}. \tag{4.67}$$

Thus we can now use the matrix fractions to solve $\mathbf{\Sigma}(\Delta t) = \mathbf{C}_W(\Delta t)\,\mathbf{D}_W^{-1}(\Delta t)$, where

$$\begin{pmatrix} \mathbf{C}_W(\Delta t) \\ \mathbf{D}_W(\Delta t) \end{pmatrix} = \exp\left\{ \begin{pmatrix} \mathbf{F} & \mathbf{L}\mathbf{Q}\mathbf{L}^\top \\ \mathbf{0} & -\mathbf{F}^\top \end{pmatrix} \Delta t \right\} \begin{pmatrix} \mathbf{0} \\ \mathbf{I} \end{pmatrix}. \tag{4.68}$$

Example 4.6 (Discretization of spring model)**.** *Recall that the spring model in Example 2.8, which in proper SDE interpretation has the form*

$$\underbrace{\begin{pmatrix} \mathrm{d}x_1(t) \\ \mathrm{d}x_2(t) \end{pmatrix}}_{\mathrm{d}\mathbf{x}(t)} = \underbrace{\begin{pmatrix} 0 & 1 \\ -\nu^2 & -\gamma \end{pmatrix}}_{\mathbf{F}} \underbrace{\begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix}}_{\mathbf{x}} \mathrm{d}t + \underbrace{\begin{pmatrix} 0 \\ 1 \end{pmatrix}}_{\mathbf{L}} \mathrm{d}\beta(t). \tag{4.69}$$

If we now wish to discretize this model, we encounter the problem that already the matrix exponential for $\mathbf{A}(\Delta t)$ is fairly complicated and we cannot hope to compute the integral for $\mathbf{\Sigma}(\Delta t)$ in Equation (4.61) *in closed form. Hence, numerical integration would be needed which can be cumbersome. However, for given values of the parameters, we can numerically use the matrix fraction decomposition and do the discretization with the following simple Matlab code snippet:*

```
nu = 1;
ga = 1/10;
dt = 1/2;

F = [0 1; -nu^2 -ga];
L = [0; 1];
```

```
Q = 1;
dim = size(F,1);

M  = [F L*Q*L'; zeros(dim) -F'];
A  = expm(F*dt)
CD = expm(M*dt)*[zeros(dim); eye(dim)];
S  = CD(1:dim,:) / CD(dim+1:end,:)
```

which outputs the matrices:

```
A =

    0.8796    0.4676
   -0.4676    0.8328

S =

    0.0382    0.1093
    0.1093    0.4386
```

This same functionality has also been implemented in the function `lti_disc.m` *in the EKF/UKF Toolbox (http://becs.aalto.fi/en/research/bayes/ekfukf/).*

Remark 4.3 (LTI SDE with constant input)**.** *Note the that by the first glance the above method does not seem to directly work for discretization of LTI SDEs with a constant input* $\mathbf{u}$*:*

$$\mathrm{d}\mathbf{x} = \mathbf{F}\,\mathbf{x}(t)\,\mathrm{d}t + \mathbf{u}\,\mathrm{d}t + \mathbf{L}\,\mathrm{d}\boldsymbol{\beta}(t), \tag{4.70}$$

but it turns out that it actually does. This is because we can rewrite the equation as

$$\begin{aligned} \mathrm{d}\mathbf{x} &= \mathbf{F}\,\mathbf{x}\,\mathrm{d}t + \mathbf{u}\,\mathrm{d}t + \mathbf{L}\,\mathrm{d}\boldsymbol{\beta}(t) \\ \mathrm{d}\mathbf{u} &= \mathbf{0}. \end{aligned} \tag{4.71}$$

The discretization can now be done to the joint state space $(\mathbf{x}, \mathbf{u})$*, which then gives one additional coefficient* $\mathbf{B}(\Delta t)$ *for the discretization:*

$$\mathbf{x}(t_{k+1}) = \mathbf{A}(\Delta t_k)\,\mathbf{x}(t_k) + \mathbf{B}(\Delta t_k)\,\mathbf{u} + \mathbf{q}_k. \tag{4.72}$$

If the input is some time-dependent $\mathbf{u}(t)$*, we can also directly use this result to form a zeroth-order-hold (ZOH) approximation to the input contribution. However, by more complicated augmentation tricks, we can also construct higher order approximations with respect to the input.*

Chapter 5

Linearization and Itô–Taylor series of SDEs

5.1 Gaussian approximations

In the previous chapter we saw that the differential equations for the mean and covariance of the solution to the SDE

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\,\mathrm{d}t + \mathbf{L}(\mathbf{x}, t)\,\mathrm{d}\boldsymbol{\beta}, \qquad \mathbf{x}(0) \sim p(\mathbf{x}(0)), \tag{5.1}$$

are

$$\frac{\mathrm{d}\mathbf{m}}{\mathrm{d}t} = \mathrm{E}\left[\mathbf{f}(\mathbf{x}, t)\right], \tag{5.2}$$

$$\begin{aligned}\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}t} &= \mathrm{E}\left[\mathbf{f}(\mathbf{x}, t)\,(\mathbf{x} - \mathbf{m})^{\mathsf{T}}\right] + \mathrm{E}\left[(\mathbf{x} - \mathbf{m})\,\mathbf{f}^{\mathsf{T}}(\mathbf{x}, t)\right] \\ &\quad + \mathrm{E}\left[\mathbf{L}(\mathbf{x}, t)\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}(\mathbf{x}, t)\right].\end{aligned} \tag{5.3}$$

If we write down the expectation integrals explicitly, these equations can be seen to have the form

$$\frac{\mathrm{d}\mathbf{m}}{\mathrm{d}t} = \int \mathbf{f}(\mathbf{x}, t)\; p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}, \tag{5.4}$$

$$\begin{aligned}\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}t} &= \int \mathbf{f}(\mathbf{x}, t)\,(\mathbf{x} - \mathbf{m})^{\mathsf{T}}\; p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x} \\ &\quad + \int (\mathbf{x} - \mathbf{m})\,\mathbf{f}^{\mathsf{T}}(\mathbf{x}, t)\; p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x} \\ &\quad + \int \mathbf{L}(\mathbf{x}, t)\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}(\mathbf{x}, t)\; p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}.\end{aligned} \tag{5.5}$$

Because $p(\mathbf{x}, t)$ is the solution of the Fokker–Planck–Kolmogorov equation (4.2), these equations cannot be solved in practice. However, one very useful class of

approximations can be obtained by replacing the FPK solution with a Gaussian approximation as follows:

$$p(\mathbf{x}, t) \approx \mathrm{N}(\mathbf{x} \mid \mathbf{m}(t), \mathbf{P}(t)), \tag{5.6}$$

where $\mathbf{m}(t)$ and $\mathbf{P}(t)$ are the mean and covariance of the state, respectively. This approximation is referred to as the *Gaussian assumed density approximation* (Kushner, 1967), because we do the computations under the assumption that the state distribution is indeed Gaussian. It is also called *Gaussian process approximation* (Archambeau and Opper, 2011). The approximation method can be written as the following algorithm.

Algorithm 5.1 (Gaussian process approximation I)**.** *A Gaussian process approximation to the SDE* (5.1) *can be obtained by integrating the following differential equations from the initial conditions* $\mathbf{m}(0) = \mathrm{E}[\mathbf{x}(0)]$ *and* $\mathbf{P}(0) = \mathrm{Cov}[\mathbf{x}(0)]$ *to the target time* t*:*

$$\begin{aligned}
\frac{\mathrm{d}\mathbf{m}}{\mathrm{d}t} &= \int \mathbf{f}(\mathbf{x}, t)\ \mathrm{N}(\mathbf{x} \mid \mathbf{m}, \mathbf{P})\, \mathrm{d}\mathbf{x}, \\
\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}t} &= \int \mathbf{f}(\mathbf{x}, t)\, (\mathbf{x} - \mathbf{m})^{\mathsf{T}}\ \mathrm{N}(\mathbf{x} \mid \mathbf{m}, \mathbf{P})\, \mathrm{d}\mathbf{x} \\
&\quad + \int (\mathbf{x} - \mathbf{m})\, \mathbf{f}^{\mathsf{T}}(\mathbf{x}, t)\ \mathrm{N}(\mathbf{x} \mid \mathbf{m}, \mathbf{P})\, \mathrm{d}\mathbf{x} \\
&\quad + \int \mathbf{L}(\mathbf{x}, t)\, \mathbf{Q}\, \mathbf{L}^{\mathsf{T}}(\mathbf{x}, t)\ \mathrm{N}(\mathbf{x} \mid \mathbf{m}, \mathbf{P})\, \mathrm{d}\mathbf{x},
\end{aligned} \tag{5.7}$$

or if we denote the Gaussian expectation as

$$\mathrm{E}_{\mathrm{N}}[\mathbf{g}(\mathbf{x})] = \int \mathbf{g}(\mathbf{x})\ \mathrm{N}(\mathbf{x} \mid \mathbf{m}, \mathbf{P})\, \mathrm{d}\mathbf{x}, \tag{5.8}$$

the equations can be written as

$$\begin{aligned}
\frac{\mathrm{d}\mathbf{m}}{\mathrm{d}t} &= \mathrm{E}_{\mathrm{N}}[\mathbf{f}(\mathbf{x}, t)], \\
\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}t} &= \mathrm{E}_{\mathrm{N}}[(\mathbf{x} - \mathbf{m})\, \mathbf{f}^{\mathsf{T}}(\mathbf{x}, t)] + \mathrm{E}_{\mathrm{N}}[\mathbf{f}(\mathbf{x}, t)\, (\mathbf{x} - \mathbf{m})^{\mathsf{T}}] \\
&\quad + \mathrm{E}_{\mathrm{N}}[\mathbf{L}(\mathbf{x}, t)\, \mathbf{Q}\, \mathbf{L}^{\mathsf{T}}(\mathbf{x}, t)],
\end{aligned} \tag{5.9}$$

If the function $\mathbf{x} \mapsto \mathbf{f}(\mathbf{x}, t)$ is differentiable, the covariance differential equation can be simplified by using the following well known property of Gaussian random variables:

Theorem 5.1. *Let* $\mathbf{f}(\mathbf{x}, t)$ *be differentiable with respect to* $\mathbf{x}$ *and let* $\mathbf{x} \sim \mathrm{N}(\mathbf{m}, \mathbf{P})$*. Then the following identity holds (see, e.g., Papoulis, 1984; Särkkä and Sarmavuori,*

2013):

$$\int \mathbf{f}(\mathbf{x},t)\,(\mathbf{x}-\mathbf{m})^{\mathsf{T}}\;\mathrm{N}(\mathbf{x}\mid\mathbf{m},\mathbf{P})\,\mathrm{d}\mathbf{x} = \left[\int \mathbf{F}_x(\mathbf{x},t)\;\mathrm{N}(\mathbf{x}\mid\mathbf{m},\mathbf{P})\,\mathrm{d}\mathbf{x}\right]\mathbf{P}, \quad (5.10)$$

where $\mathbf{F}_x(\mathbf{x},t)$ *is the Jacobian matrix of* $\mathbf{f}(\mathbf{x},t)$ *with respect to* $\mathbf{x}$.

Using the theorem, the mean and covariance Equations (5.9) can be equivalently written as follows.

Algorithm 5.2 (Gaussian process approximation II)**.** *A Gaussian process approximation to the SDE* (5.1) *can be obtained by integrating the following differential equations from the initial conditions* $\mathbf{m}(0) = \mathrm{E}[\mathbf{x}(0)]$ *and* $\mathbf{P}(0) = \mathrm{Cov}[\mathbf{x}(0)]$ *to the target time* t*:*

$$\begin{aligned}\frac{\mathrm{d}\mathbf{m}}{\mathrm{d}t} &= \mathrm{E}_{\mathrm{N}}[\mathbf{f}(\mathbf{x},t)],\\ \frac{\mathrm{d}\mathbf{P}}{\mathrm{d}t} &= \mathbf{P}\;\mathrm{E}_{\mathrm{N}}[\mathbf{F}_x(\mathbf{x},t)]^{\mathsf{T}} + \mathrm{E}_{\mathrm{N}}[\mathbf{F}_x(\mathbf{x},t)]\,\mathbf{P} + \mathrm{E}_{\mathrm{N}}[\mathbf{L}(\mathbf{x},t)\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}(\mathbf{x},t)],\end{aligned} \quad (5.11)$$

where $\mathrm{E}_{\mathrm{N}}[\cdot]$ *denotes the expectation with respect to* $\mathbf{x} \sim \mathrm{N}(\mathbf{m},\mathbf{P})$.

The approximations presented in this section are formally equivalent to so called statistical linearization approximations (Gelb, 1974; Socha, 2008) and they are also closely related to the variational approximations of Archambeau and Opper (2011).

5.2 Linearization and sigma-point approximations

In the previous section we presented how to form Gaussian approximations of SDEs. However, to implement the method one is required to compute the following kind of n-dimensional Gaussian integrals:

$$\mathrm{E}_{\mathrm{N}}[\mathbf{g}(\mathbf{x},t)] = \int \mathbf{g}(\mathbf{x},t)\;\mathrm{N}(\mathbf{x}\mid\mathbf{m},\mathbf{P})\,\mathrm{d}\mathbf{x}. \quad (5.12)$$

A classical approach which is very common in filtering theory (Jazwinski, 1970; Maybeck, 1982) is to linearize the drift $\mathbf{f}(\mathbf{x},t)$ around the mean $\mathbf{m}$ as follows:

$$\mathbf{f}(\mathbf{x},t) \approx \mathbf{f}(\mathbf{m},t) + \mathbf{F}_x(\mathbf{m},t)\,(\mathbf{x}-\mathbf{m}), \quad (5.13)$$

and to approximate the expectation of the diffusion part as

$$\mathbf{L}(\mathbf{x},t) \approx \mathbf{L}(\mathbf{m},t). \quad (5.14)$$

This leads to the following approximation, which is commonly used in extended Kalman filters (EKF).

Algorithm 5.3 (Linearization approximation of SDE)**.** *A linearization based approximation to the SDE* (5.1) *can be obtained by integrating the following differential equations from the initial conditions* $\mathbf{m}(0) = \mathrm{E}[\mathbf{x}(0)]$ *and* $\mathbf{P}(0) = \mathrm{Cov}[\mathbf{x}(0)]$ *to the target time* t*:*

$$\begin{aligned} \frac{\mathrm{d}\mathbf{m}}{\mathrm{d}t} &= \mathbf{f}(\mathbf{m}, t), \\ \frac{\mathrm{d}\mathbf{P}}{\mathrm{d}t} &= \mathbf{P}\,\mathbf{F}_x^{\mathsf{T}}(\mathbf{m}, t) + \mathbf{F}_x(\mathbf{m}, t)\,\mathbf{P} + \mathbf{L}(\mathbf{m}, t)\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}(\mathbf{m}, t). \end{aligned} \tag{5.15}$$

Another general class of approximations are the Gauss–Hermite cubature type of approximations where we approximate the integrals as weighted sums:

$$\int \mathbf{f}(\mathbf{x}, t)\ \mathrm{N}(\mathbf{x} \mid \mathbf{m}, \mathbf{P})\,\mathrm{d}\mathbf{x} \approx \sum_i W^{(i)}\,\mathbf{f}(\mathbf{x}^{(i)}, t), \tag{5.16}$$

where $\mathbf{x}^{(i)}$ and $W^{(i)}$ are the sigma points (abscissas) and their accompanying weights, which have been selected using a method specific deterministic rule. This kind of rules are commonly used in the context of filtering theory (*cf.* Särkkä and Sarmavuori, 2013). In multidimensional Gauss–Hermite integration, the unscented transform, and cubature integration, the sigma points are selected as follows:

$$\mathbf{x}^{(i)} = \mathbf{m} + \sqrt{\mathbf{P}}\,\boldsymbol{\xi}_i, \tag{5.17}$$

where the matrix square root is defined by $\mathbf{P} = \sqrt{\mathbf{P}}\,\sqrt{\mathbf{P}}^{\mathsf{T}}$ (typically Cholesky factorization), and the points $\boldsymbol{\xi}_i$ and the weights $W^{(i)}$ are selected as follows:

The Gauss–Hermite integration method (the product rule based method) uses a set of m^n vectors $\boldsymbol{\xi}_i$, which have been formed as a Cartesian product of zeros of the Hermite polynomials of order m. The weights $W^{(i)}$ are formed as products of the corresponding one-dimensional Gauss–Hermite integration weights (see, Ito and Xiong, 2000; Wu et al., 2006, for details).

The Unscented transform uses a zero vector (origin) and $2n$ unit coordinate vectors $\mathbf{e}_i$ as follows (the method can also be generalized a bit):

$$\begin{aligned} \boldsymbol{\xi}_0 &= 0, \\ \boldsymbol{\xi}_i &= \begin{cases} \sqrt{\lambda + n}\,\mathbf{e}_i, & i = 1, \ldots, n, \\ -\sqrt{\lambda + n}\,\mathbf{e}_{i-n}, & i = n+1, \ldots, 2n, \end{cases} \end{aligned} \tag{5.18}$$

and the weights are defined as follows:

$$\begin{aligned} W^{(0)} &= \frac{\lambda}{n + \kappa}, \\ W^{(i)} &= \frac{1}{2(n + \kappa)}, \qquad i = 1, \ldots, 2n, \end{aligned} \tag{5.19}$$

where κ is a parameter of the method.

The Cubature method (spherical 3rd degree) uses only $2n$ vectors as follows:

$$\boldsymbol{\xi}_i = \begin{cases} \sqrt{n}\,\mathbf{e}_i, & i = 1,\ldots,n, \\ -\sqrt{n}\,\mathbf{e}_{i-n}, & i = n+1,\ldots,2n, \end{cases} \tag{5.20}$$

and the weights are defined as $W^{(i)} = 1/(2n)$, for $i = 1, 2, \ldots, 2n$.

The sigma point methods above lead to the following approximations to the prediction differential equations.

Algorithm 5.4 (Sigma-point approximation of SDEs)**.** *A sigma-point based approximation to the SDE* (5.1) *can be obtained by integrating the following differential equations from the initial conditions* $\mathbf{m}(0) = \mathrm{E}[\mathbf{x}(0)]$ *and* $\mathbf{P}(0) = \mathrm{Cov}[\mathbf{x}(0)]$ *to the target time* t*:*

$$\begin{aligned} \frac{\mathrm{d}\mathbf{m}}{\mathrm{d}t} &= \sum_i W^{(i)}\,\mathbf{f}(\mathbf{m} + \sqrt{\mathbf{P}}\,\boldsymbol{\xi}_i, t), \\ \frac{\mathrm{d}\mathbf{P}}{\mathrm{d}t} &= \sum_i W^{(i)}\,\mathbf{f}(\mathbf{m} + \sqrt{\mathbf{P}}\,\boldsymbol{\xi}_i, t)\,\boldsymbol{\xi}_i^{\mathsf{T}}\,\sqrt{\mathbf{P}}^{\mathsf{T}} \\ &\quad + \sum_i W^{(i)}\,\sqrt{\mathbf{P}}\,\boldsymbol{\xi}_i\,\mathbf{f}^{\mathsf{T}}(\mathbf{m} + \sqrt{\mathbf{P}}\,\boldsymbol{\xi}_i, t) \\ &\quad + \sum_i W^{(i)}\,\mathbf{L}(\mathbf{m} + \sqrt{\mathbf{P}}\,\boldsymbol{\xi}_i, t)\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}(\mathbf{m} + \sqrt{\mathbf{P}}\,\boldsymbol{\xi}_i, t). \end{aligned} \tag{5.21}$$

Once the Gaussian integral approximation has been selected, the solutions to the resulting ordinary differential equations can be computed, for example, by the fourth order Runge–Kutta method or some similar numerical ODE solution method. It would also be possible to approximate the integrals using various other methods from filtering theory (see, *e.g.*, Jazwinski, 1970; Wu et al., 2006; Särkkä and Sarmavuori, 2013).

Example 5.1. *Consider the non-linear Itô stochastic differential equation model*

$$\mathrm{d}x = -\left(\frac{1}{10}\right)^2 \sin(x)\,\cos^3(x)\,\mathrm{d}t + \frac{1}{10}\cos^2(x)\,\mathrm{d}\beta, \qquad x(0) = x_0. \tag{5.22}$$

This model has the solution $x(t) = \arctan(1/10\,\beta(t) + \tan(x_0))$ *which we will use as ground truth, where* $\beta(t)$ *is a standard Brownian motion. In this example, let* $x_0 = 1$*. The goal is to characterize the solution at* $t = 10$ *using a Gaussian approximation* $\tilde{x}(t) \approx \mathrm{N}(m(t), P(t))$ *of the exact solution.*

From the model we have the drift $f(x) = -(1/10)^2 \sin(x)\,\cos^3(x)$ *and diffusion* $L(x) = 1/10\,\cos^2(x)$*. We use a the Cubature integration sigma-point scheme, which gives us the following two-dimensional ordinary differential equation model for the mean and covariance:* $\dot{\mathbf{z}}(t) = (\mathrm{d}m(t)/\mathrm{d}t\ ,\ \mathrm{d}P(t)/\mathrm{d}t\,)$*. Applying the Cubature formula (*$\xi = \pm 1, W^{(i)} = 1/2\ , i = 1, 2$*) gives the following*

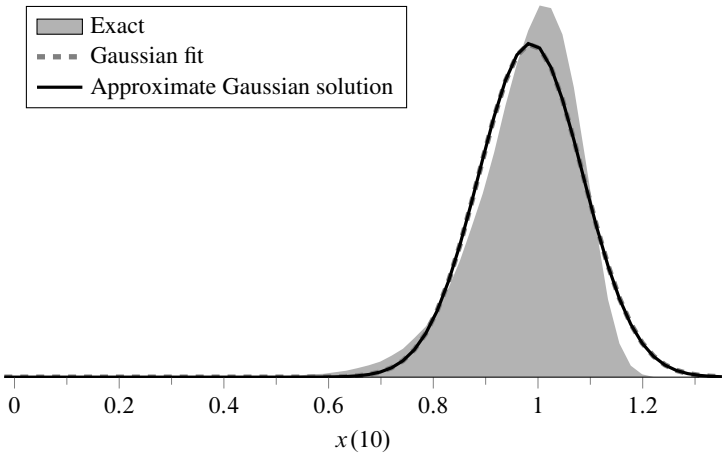


Figure 5.1: An example of a sigma-point based Gaussian approximation to a non-linear SDE. The exact solution at $t = 10$ is shown by the patch, and the dashed line illustrates the Gaussian fit to it. The approximation is shown in solid black.

integrand:

$$\dot{z}_1 = \frac{1}{2} f(z_1 - \sqrt{z_2}) + \frac{1}{2} f(z_1 + \sqrt{z_2}), \tag{5.23}$$

$$\dot{z}_2 = \sqrt{z_2}\, f(z_1 - \sqrt{z_2}) - \sqrt{z_2}\, f(z_1 + \sqrt{z_2}) + \frac{1}{2}\left[L(z_1 - \sqrt{z_2})\right]^2 + \frac{1}{2}\left[L(z_1 + \sqrt{z_2})\right]^2, \tag{5.24}$$

where $\mathbf{z}(0) = (x_0, 0)$. *We use the fourth order Runge–Kutta scheme for solving* $\mathbf{z}(10)$ *with a step size of* $\Delta t = 2^{-6}$. *Figure 5.1 illustrates the exact solution of* $x(10)$ *by a shaded patch, and shows the Gaussian fit to it by a dashed line. The ODE-based approximative solution* $\tilde{x}(10) = \mathrm{N}(m(10), P(10))$ *is shown in solid black, and it coincides well with the Gaussian fit.*

5.3 Taylor series of ODEs

One way to find approximate solutions of deterministic ordinary differential equations (ODEs) is by using Taylor series expansions (in time direction). Even though this method as a practical ODE numerical approximation method is quite much superseded by Runge–Kutta type of derivative-free methods, it still is an important theoretical tool for finding and analyzing numerical schemes (*e.g.*, the theory of Runge–Kutta methods is based on Taylor series). In the case of SDEs, the corresponding Itô–Taylor series solutions provide a useful basis for numerical methods for SDEs. However, we cannot simply apply the same ODE-based numerical schemes to SDEs. This is because of the inconvenient fact that in the stochastic case, Runge–Kutta methods are not as easy to use as in the case of ODEs.

In this section, we derive the Taylor series based solutions of ODEs in detail, because the derivation of the Itô–Taylor series can be done in an analogous way. As the idea is the same, by first going through the deterministic case it is easy to see the essential things behind the technical details also in the SDE case.

Let's start by considering the following differential equation:

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}(t), t), \qquad \mathbf{x}(t_0) = \text{given}, \tag{5.25}$$

which can be integrated to give

$$\mathbf{x}(t) = \mathbf{x}(t_0) + \int_{t_0}^{t} \mathbf{f}(\mathbf{x}(\tau), \tau)\,\mathrm{d}\tau. \tag{5.26}$$

If the function $\mathbf{f}$ is differentiable, we can also write $t \mapsto \mathbf{f}(\mathbf{x}(t), t)$ as the solution to the differential equation

$$\frac{\mathrm{d}\mathbf{f}(\mathbf{x}(t), t)}{\mathrm{d}t} = \frac{\partial}{\partial t}\mathbf{f}(\mathbf{x}(t), t) + \sum_i f_i(\mathbf{x}(t), t)\,\frac{\partial}{\partial x_i}\mathbf{f}(\mathbf{x}(t), t), \tag{5.27}$$

where $\mathbf{f}(\mathbf{x}(t_0), t_0)$ is the given initial condition. The integral form of this is

$$\mathbf{f}(\mathbf{x}(t), t) = \mathbf{f}(\mathbf{x}(t_0), t_0) + \int_{t_0}^{t} \left[\frac{\partial}{\partial t}\mathbf{f}(\mathbf{x}(\tau), \tau) + \sum_i f_i(\mathbf{x}(\tau), \tau)\,\frac{\partial}{\partial x_i}\mathbf{f}(\mathbf{x}(\tau), \tau) \right] \mathrm{d}\tau. \tag{5.28}$$

At this point it is convenient to define the linear operator

$$\mathscr{L}(\bullet) = \frac{\partial}{\partial t}(\bullet) + \sum_i f_i\,\frac{\partial}{\partial x_i}(\bullet), \tag{5.29}$$

and rewrite the integral equation as

$$\mathbf{f}(\mathbf{x}(t), t) = \mathbf{f}(\mathbf{x}(t_0), t_0) + \int_{t_0}^{t} \mathscr{L}\,\mathbf{f}(\mathbf{x}(\tau), \tau)\,\mathrm{d}\tau. \tag{5.30}$$

Substituting this into Equation (5.26) gives

$$\begin{aligned} \mathbf{x}(t) &= \mathbf{x}(t_0) + \int_{t_0}^{t} [\mathbf{f}(\mathbf{x}(t_0), t_0) + \int_{t_0}^{\tau} \mathscr{L}\,\mathbf{f}(\mathbf{x}(\tau), \tau)\,\mathrm{d}\tau]\,\mathrm{d}\tau \\ &= \mathbf{x}(t_0) + \mathbf{f}(\mathbf{x}(t_0), t_0)\,(t - t_0) + \int_{t_0}^{t}\int_{t_0}^{\tau} \mathscr{L}\,\mathbf{f}(\mathbf{x}(\tau), \tau)\,\mathrm{d}\tau\,\mathrm{d}\tau. \end{aligned} \tag{5.31}$$

The term in the integrand on the right can again be defined as solution to the differential equation

$$\begin{aligned} \frac{\mathrm{d}[\mathscr{L}\,\mathbf{f}(\mathbf{x}(t), t)]}{\mathrm{d}t} &= \frac{\partial[\mathscr{L}\,\mathbf{f}(\mathbf{x}(t), t)]}{\partial t} + \sum_i f_i(\mathbf{x}(t), t)\,\frac{\partial[\mathscr{L}\,\mathbf{f}(\mathbf{x}(t), t)]}{\partial x_i} \\ &= \mathscr{L}^2\,\mathbf{f}(\mathbf{x}(t), t). \end{aligned} \tag{5.32}$$

which in integral form is

$$\mathscr{L}\,\mathbf{f}\,(\mathbf{x}(t),t) = \mathscr{L}\,\mathbf{f}\,(\mathbf{x}(t_0),t_0) + \int_{t_0}^{t} \mathscr{L}^2\,\mathbf{f}\,(\mathbf{x}(\tau),\tau)\,\mathrm{d}\tau. \tag{5.33}$$

Substituting into the equation of $\mathbf{x}(t)$ then gives

$$\begin{aligned}\mathbf{x}(t) &= \mathbf{x}(t_0) + \mathbf{f}\,(\mathbf{x}(t_0),t)\,(t-t_0)\\ &\quad + \int_{t_0}^{t}\int_{t_0}^{\tau} [\mathscr{L}\,\mathbf{f}\,(\mathbf{x}(t_0),t_0) + \int_{t_0}^{\tau} \mathscr{L}^2\,\mathbf{f}\,(\mathbf{x}(\tau),\tau)\,\mathrm{d}\tau]\,\mathrm{d}\tau\,\mathrm{d}\tau\\ &= \mathbf{x}(t_0) + \mathbf{f}\,(\mathbf{x}(t_0),t_0)\,(t-t_0) + \frac{1}{2}\,\mathscr{L}\,\mathbf{f}\,(\mathbf{x}(t_0),t_0)\,(t-t_0)^2\\ &\quad + \int_{t_0}^{t}\int_{t_0}^{\tau}\int_{t_0}^{\tau} \mathscr{L}^2\,\mathbf{f}\,(\mathbf{x}(\tau),\tau)\,\mathrm{d}\tau\,\mathrm{d}\tau\,\mathrm{d}\tau.\end{aligned} \tag{5.34}$$

If we continue this procedure *ad infinitum*, we obtain the following *Taylor series expansion* for the solution of the ODE:

$$\begin{aligned}\mathbf{x}(t) &= \mathbf{x}(t_0) + \mathbf{f}\,(\mathbf{x}(t_0),t_0)\,(t-t_0) + \frac{1}{2!}\,\mathscr{L}\,\mathbf{f}\,(\mathbf{x}(t_0),t_0)\,(t-t_0)^2\\ &\quad + \frac{1}{3!}\,\mathscr{L}^2\,\mathbf{f}\,(\mathbf{x}(t_0),t_0)\,(t-t_0)^3 + \ldots\end{aligned} \tag{5.35}$$

From the above derivation we also get the result that if we truncate the series at the nth term, the residual error is

$$\mathbf{r}_n(t) = \int_{t_0}^{t} \cdots \int_{t_0}^{\tau} \mathscr{L}^n\,\mathbf{f}\,(\mathbf{x}(\tau),\tau)\,\mathrm{d}\tau^{n+1}, \tag{5.36}$$

which could be further simplified via integration by parts and using the mean value theorem. To derive the series expansion for an arbitrary function $\mathbf{x}(t)$, we can define it as solution to the trivial differential equation

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{f}\,(t), \qquad \mathbf{x}(t_0) = \text{given}, \tag{5.37}$$

where $\mathbf{f}\,(t) = \mathrm{d}\mathbf{x}(t)/\mathrm{d}t$. Because $\mathbf{f}$ is independent of $\mathbf{x}$, we have

$$\mathscr{L}^n\mathbf{f} = \frac{\mathrm{d}^{n+1}\mathbf{x}(t)}{\mathrm{d}t^{n+1}}. \tag{5.38}$$

Thus the corresponding series becomes the classical Taylor series:

$$\begin{aligned}\mathbf{x}(t) &= \mathbf{x}(t_0) + \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}(t_0)\,(t-t_0) + \frac{1}{2!}\,\frac{\mathrm{d}^2\mathbf{x}}{\mathrm{d}t^2}(t_0)\,(t-t_0)^2\\ &\quad + \frac{1}{3!}\,\frac{\mathrm{d}^3\mathbf{x}}{\mathrm{d}t^3}(t_0)\,(t-t_0)^3 + \ldots\end{aligned} \tag{5.39}$$

5.4 Itô–Taylor series based strong approximations of SDEs

Itô–Taylor series (see Kloeden et al., 1994; Kloeden and Platen, 1999) is an extension of the Taylor series of ODEs to SDEs. The derivation is identical to the Taylor series solution in the previous section except that we replace the time derivative computations with application of the Itô formula.

Let's consider the following SDE

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}(t), t)\,\mathrm{d}t + \mathbf{L}(\mathbf{x}(t), t)\,\mathrm{d}\boldsymbol{\beta}, \qquad \mathbf{x}(t_0) \sim p(\mathbf{x}(t_0)). \tag{5.40}$$

In integral form this equation can be expressed as

$$\mathbf{x}(t) = \mathbf{x}(t_0) + \int_{t_0}^{t} \mathbf{f}(\mathbf{x}(\tau), \tau)\,\mathrm{d}\tau + \int_{t_0}^{t} \mathbf{L}(\mathbf{x}(\tau), \tau)\,\mathrm{d}\boldsymbol{\beta}(\tau). \tag{5.41}$$

Applying the Itô formula to terms $\mathbf{f}(\mathbf{x}(t), t)$ and $\mathbf{L}(\mathbf{x}(t), t)$ gives the following for the drift

$$\begin{aligned}
\mathrm{d}\mathbf{f}(\mathbf{x}(t), t) = {} & \frac{\partial \mathbf{f}(\mathbf{x}(t), t)}{\partial t}\,\mathrm{d}t + \sum_u \frac{\partial \mathbf{f}(\mathbf{x}(t), t)}{\partial x_u}\, f_u(\mathbf{x}(t), t)\,\mathrm{d}t \\
& + \sum_u \frac{\partial \mathbf{f}(\mathbf{x}(t), t)}{\partial x_u}\,[\mathbf{L}(\mathbf{x}(t), t)\,\mathrm{d}\boldsymbol{\beta}(\tau)]_u \\
& + \frac{1}{2}\sum_{uv} \frac{\partial^2 \mathbf{f}(\mathbf{x}(t), t)}{\partial x_u\,\partial x_v}\,[\mathbf{L}(\mathbf{x}(t), t)\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}(\mathbf{x}(t), t)]_{uv}\,\mathrm{d}t,
\end{aligned} \tag{5.42}$$

and the following for the diffusion

$$\begin{aligned}
\mathrm{d}\mathbf{L}(\mathbf{x}(t), t) = {} & \frac{\partial \mathbf{L}(\mathbf{x}(t), t)}{\partial t}\,\mathrm{d}t + \sum_u \frac{\partial \mathbf{L}(\mathbf{x}(t), t)}{\partial x_u}\, f_u(\mathbf{x}(t), t)\,\mathrm{d}t \\
& + \sum_u \frac{\partial \mathbf{L}(\mathbf{x}(t), t)}{\partial x_u}\,[\mathbf{L}(\mathbf{x}(t), t)\,\mathrm{d}\boldsymbol{\beta}(\tau)]_u \\
& + \frac{1}{2}\sum_{uv} \frac{\partial^2 \mathbf{L}(\mathbf{x}(t), t)}{\partial x_u\,\partial x_v}\,[\mathbf{L}(\mathbf{x}(t), t)\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}(\mathbf{x}(t), t)]_{uv}\,\mathrm{d}t.
\end{aligned} \tag{5.43}$$

In integral form these can be written as

$$\begin{aligned}
\mathbf{f}(\mathbf{x}(t), t) = {} & \mathbf{f}(\mathbf{x}(t_0), t_0) + \int_{t_0}^{t} \frac{\partial \mathbf{f}(\mathbf{x}(\tau), \tau)}{\partial t}\,\mathrm{d}\tau \\
& + \int_{t_0}^{t} \sum_u \frac{\partial \mathbf{f}(\mathbf{x}(\tau), \tau)}{\partial x_u}\, f_u(\mathbf{x}(\tau), \tau)\,\mathrm{d}\tau \\
& + \int_{t_0}^{t} \sum_u \frac{\partial \mathbf{f}(\mathbf{x}(\tau), \tau)}{\partial x_u}\,[\mathbf{L}(\mathbf{x}(\tau), \tau)\,\mathrm{d}\boldsymbol{\beta}(\tau)]_u \\
& + \int_{t_0}^{t} \frac{1}{2}\sum_{uv} \frac{\partial^2 \mathbf{f}(\mathbf{x}(\tau), \tau)}{\partial x_u\,\partial x_v}\,[\mathbf{L}(\mathbf{x}(\tau), \tau)\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}(\mathbf{x}(\tau), \tau)]_{uv}\,\mathrm{d}\tau,
\end{aligned} \tag{5.44}$$

and

$$\begin{aligned}\mathbf{L}(\mathbf{x}(t),t) = \mathbf{L}(\mathbf{x}(t_0),t_0) &+ \int_{t_0}^{t} \frac{\partial \mathbf{L}(\mathbf{x}(\tau),\tau)}{\partial t}\,\mathrm{d}\tau \\ &+ \int_{t_0}^{t} \sum_u \frac{\partial \mathbf{L}(\mathbf{x}(\tau),\tau)}{\partial x_u}\, f_u(\mathbf{x}(\tau),\tau)\,\mathrm{d}\tau \\ &+ \int_{t_0}^{t} \sum_u \frac{\partial \mathbf{L}(\mathbf{x}(\tau),\tau)}{\partial x_u}\, [\mathbf{L}(\mathbf{x}(\tau),\tau)\,\mathrm{d}\boldsymbol{\beta}(\tau)]_u \\ &+ \int_{t_0}^{t} \frac{1}{2} \sum_{uv} \frac{\partial^2 \mathbf{L}(\mathbf{x}(\tau),\tau)}{\partial x_u\,\partial x_v}\, [\mathbf{L}(\mathbf{x}(\tau),\tau)\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}(\mathbf{x}(\tau),\tau)]_{uv}\,\mathrm{d}\tau.\end{aligned} \tag{5.45}$$

If we define the following two operators

$$\begin{aligned}\mathscr{L}_t\,(\bullet) &= \frac{\partial(\bullet)}{\partial t} + \sum_u \frac{\partial(\bullet)}{\partial x_u}\, f_u + \frac{1}{2} \sum_{uv} \frac{\partial^2(\bullet)}{\partial x_u\,\partial x_v}\, [\mathbf{L}\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}]_{uv} \\ \mathscr{L}_{\beta,v}\,(\bullet) &= \sum_u \frac{\partial(\bullet)}{\partial x_u}\, \mathbf{L}_{uv}, \qquad \text{for } v = 1,\ldots,n,\end{aligned} \tag{5.46}$$

then we can conveniently write

$$\begin{aligned}\mathbf{f}(\mathbf{x}(t),t) &= \mathbf{f}(\mathbf{x}(t_0),t_0) + \int_{t_0}^{t} \mathscr{L}_t \mathbf{f}(\mathbf{x}(\tau),\tau)\,\mathrm{d}\tau \\ &\quad + \sum_v \int_{t_0}^{t} \mathscr{L}_{\beta,v} \mathbf{f}(\mathbf{x}(\tau),\tau)\,\mathrm{d}\beta_v(\tau), \\ \mathbf{L}(\mathbf{x}(t),t) &= \mathbf{L}(\mathbf{x}(t_0),t_0) + \int_{t_0}^{t} \mathscr{L}_t \mathbf{L}(\mathbf{x}(\tau),\tau)\,\mathrm{d}\tau \\ &\quad + \sum_v \int_{t_0}^{t} \mathscr{L}_{\beta,v} \mathbf{L}(\mathbf{x}(\tau),\tau)\,\mathrm{d}\beta_v(\tau).\end{aligned} \tag{5.47}$$

If we now substitute these into the expression of $\mathbf{x}(t)$ in Equation (6.22), we get

$$\begin{aligned}\mathbf{x}(t) = \mathbf{x}(t_0) &+ \mathbf{f}(\mathbf{x}(t_0),t_0)\,(t-t_0) + \mathbf{L}(\mathbf{x}(t_0),t_0)\,(\boldsymbol{\beta}(t)-\boldsymbol{\beta}(t_0)) \\ &+ \int_{t_0}^{t}\int_{t_0}^{\tau} \mathscr{L}_t \mathbf{f}(\mathbf{x}(\tau),\tau)\,\mathrm{d}\tau\,\mathrm{d}\tau + \sum_v \int_{t_0}^{t}\int_{t_0}^{\tau} \mathscr{L}_{\beta,v} \mathbf{f}(\mathbf{x}(\tau),\tau)\;\mathrm{d}\beta_v(\tau)\,\mathrm{d}\tau \\ &+ \int_{t_0}^{t}\int_{t_0}^{\tau} \mathscr{L}_t \mathbf{L}(\mathbf{x}(\tau),\tau)\;\mathrm{d}\tau\,\mathrm{d}\boldsymbol{\beta}(\tau) \\ &+ \sum_v \int_{t_0}^{t}\int_{t_0}^{\tau} \mathscr{L}_{\beta,v} \mathbf{L}(\mathbf{x}(\tau),\tau)\;\mathrm{d}\beta_v(\tau)\,\mathrm{d}\boldsymbol{\beta}(\tau).\end{aligned} \tag{5.48}$$

This can be seen to have the form

$$\mathbf{x}(t) = \mathbf{x}(t_0) + \mathbf{f}(\mathbf{x}(t_0),t_0)\,(t-t_0) + \mathbf{L}(\mathbf{x}(t_0),t_0)\,(\boldsymbol{\beta}(t)-\boldsymbol{\beta}(t_0)) + \mathbf{r}(t), \tag{5.49}$$

where the remainder $\mathbf{r}(t)$ consists of higher order multiple stochastic integrals involving the function itself, the drift and diffusion, and their derivatives such that

$$\begin{aligned}\mathbf{r}(t) = \int_{t_0}^{t}\int_{t_0}^{\tau} \mathcal{L}_t \mathbf{f}(\mathbf{x}(\tau),\tau)\,\mathrm{d}\tau\,\mathrm{d}\tau + \sum_v \int_{t_0}^{t}\int_{t_0}^{\tau} \mathcal{L}_{\beta,v}\mathbf{f}(\mathbf{x}(t),t)\,\mathrm{d}\beta_v(\tau)\,\mathrm{d}\tau \\ + \int_{t_0}^{t}\int_{t_0}^{\tau} \mathcal{L}_t \mathbf{L}(\mathbf{x}(\tau),\tau)\,\mathrm{d}\tau\,\mathrm{d}\boldsymbol{\beta}(\tau) \\ + \sum_v \int_{t_0}^{t}\int_{t_0}^{\tau} \mathcal{L}_{\beta,v}\mathbf{L}(\mathbf{x}(\tau),\tau)\,\mathrm{d}\beta_v(\tau)\,\mathrm{d}\boldsymbol{\beta}(\tau). \quad (5.50)\end{aligned}$$

We can now form a first order approximation to the solution by discarding the remainder term:

$$\mathbf{x}(t) \approx \mathbf{x}(t_0) + \mathbf{f}(\mathbf{x}(t_0),t_0)\,(t-t_0) + \mathbf{L}(\mathbf{x}(t_0),t_0)\,(\boldsymbol{\beta}(t)-\boldsymbol{\beta}(t_0)). \quad (5.51)$$

This leads to the Euler–Maruyama method already discussed in Section 2.4.

Algorithm 5.5 (Euler–Maruyama method)**.** *Draw* $\hat{\mathbf{x}}_0 \sim p(\mathbf{x}_0)$ *and divide time* $[0,t]$ *interval into* K *steps of length* Δt*. At each step* k *do the following:*

1. *Draw random variable* $\Delta\boldsymbol{\beta}_k$ *from the distribution (where* $t_k = k\,\Delta t$*)*

$$\Delta\boldsymbol{\beta}_k \sim \mathrm{N}(\mathbf{0}, \mathbf{Q}\,\Delta t). \quad (5.52)$$

2. *Compute*

$$\hat{\mathbf{x}}(t_{k+1}) = \hat{\mathbf{x}}(t_k) + \mathbf{f}(\hat{\mathbf{x}}(t_k),t_k)\,\Delta t + \mathbf{L}(\hat{\mathbf{x}}(t_k),t_k)\,\Delta\boldsymbol{\beta}_k. \quad (5.53)$$

The *strong order of convergence* of a stochastic numerical integration method can be roughly defined to be the smallest exponent γ such that if we numerically solve an SDE using $n = 1/\Delta t$ steps of length Δ, then there exists a constant K such that

$$\mathrm{E}\left[|\mathbf{x}(t_n) - \hat{\mathbf{x}}(t_n)|\right] \leq K\,\Delta t^{\gamma}. \quad (5.54)$$

For stochastic methods, there also exist a second type of convergence, namely *weak order of convergence*. This will be discussed in more detail in the next section.

It can be shown (Kloeden and Platen, 1999) that in the case of the Euler–Maruyama method above (under assumptions of sufficient regularity), the strong order of convergence is $\gamma = 1/2$. However, as will be shown later on, it has the weak order of convergence $\alpha = 1$. The reason why the strong order of convergence is just $1/2$ is that the term with $\mathrm{d}\beta_v(\tau)\,\mathrm{d}\boldsymbol{\beta}(\tau)$ in the residual, when integrated, leaves us with a term with $\mathrm{d}\boldsymbol{\beta}(\tau)$ which is only of order $\mathrm{d}t^{1/2}$. Thus we can increase the strong order to one by expanding that term.

We can now do the same kind of expansion for the term $\mathscr{L}_{\beta,v}\mathbf{L}(\mathbf{x}(\tau),\tau)$ as we did in Equation (5.47), which leads to

$$\begin{aligned}\mathscr{L}_{\beta,v}\mathbf{L}(\mathbf{x}(t),t) = \mathscr{L}_{\beta,v}\mathbf{L}(\mathbf{x}(t_0),t_0) + \int_{t_0}^{t} \mathscr{L}_t\mathscr{L}_{\beta,v}\mathbf{L}(\mathbf{x}(t),t)\,\mathrm{d}t \\ + \sum_v \int_{t_0}^{t} \mathscr{L}^2_{\beta,v}\mathbf{L}(\mathbf{x}(t),t)\,\mathrm{d}\beta_v(\tau).\end{aligned} \tag{5.55}$$

Substituting this into the Equation (5.48) gives

$$\begin{aligned}\mathbf{x}(t) = \mathbf{x}(t_0) + \mathbf{f}(\mathbf{x}(t_0),t_0)\,(t-t_0) + \mathbf{L}(\mathbf{x}(t_0),t_0)\,(\boldsymbol{\beta}(t)-\boldsymbol{\beta}(t_0)) \\ + \sum_v \mathscr{L}_{\beta,v}\mathbf{L}(\mathbf{x}(t_0),t_0) \int_{t_0}^{t}\int_{t_0}^{\tau} \mathrm{d}\beta_v(\tau)\,\mathrm{d}\boldsymbol{\beta}(\tau) + \text{remainder}.\end{aligned} \tag{5.56}$$

Now the important thing is to notice the *iterated Itô integral* appearing in the equation:

$$\int_{t_0}^{t}\int_{t_0}^{\tau} \mathrm{d}\beta_v(\tau)\,\mathrm{d}\boldsymbol{\beta}(\tau). \tag{5.57}$$

Computation of this kind of integrals and more general iterated Itô integrals turns out to be quite non-trivial. However, assuming that we can indeed compute the integral, as well as draw the corresponding Brownian increment (recall that the terms are not independent), then we can form the following scheme known as *Milstein's method*.

Algorithm 5.6 (Milstein's method)**.** *Draw* $\hat{\mathbf{x}}_0 \sim p(\mathbf{x}_0)$ *and divide the time interval* $[0,t]$ *into* K *steps of length* Δt*. At each step* k *do the following:*

1. *Jointly draw a Brownian motion increment and the iterated Itô integral of it:*

$$\begin{aligned}\Delta\boldsymbol{\beta}_k &= \boldsymbol{\beta}(t_{k+1}) - \boldsymbol{\beta}(t_k) \\ \Delta\boldsymbol{\chi}_{v,k} &= \int_{t_k}^{t_{k+1}}\int_{t_k}^{\tau} \mathrm{d}\beta_v(\tau)\,\mathrm{d}\boldsymbol{\beta}(\tau).\end{aligned} \tag{5.58}$$

2. *Compute*

$$\begin{aligned}\hat{\mathbf{x}}(t_{k+1}) = \hat{\mathbf{x}}(t_k) + \mathbf{f}(\hat{\mathbf{x}}(t_k),t_k)\,\Delta t + \mathbf{L}(\hat{\mathbf{x}}(t_k),t_k)\,\Delta\boldsymbol{\beta}_k \\ + \sum_v \left[\sum_u \frac{\partial \mathbf{L}}{\partial x_u}(\hat{\mathbf{x}}(t_k),t_k)\,\mathbf{L}_{uv}(\hat{\mathbf{x}}(t_k),t_k)\right] \Delta\boldsymbol{\chi}_{v,k}.\end{aligned} \tag{5.59}$$

The strong and weak orders of the above method are both one ($\gamma = \alpha = 1$). However, the difficulty is that drawing the iterated stochastic integral jointly with the Brownian motion is hard (*cf.* Kloeden and Platen, 1999). But if the noise is additive, that is, $\mathbf{L}(\mathbf{x},t) = \mathbf{L}(t)$ then Milstein's algorithm reduces to the Euler–Maruyama method. Thus in the additive noise case the strong order of Euler–Maruyama is $\gamma = 1$ as well.

In the scalar case we can compute the iterated stochastic integral:

$$\int_{t_0}^{t}\int_{t_0}^{\tau} \mathrm{d}\beta(\tau)\,\mathrm{d}\beta(\tau) = \frac{1}{2}\left[(\beta(t)-\beta(t_0))^2 - q\,(t-t_0)\right]. \tag{5.60}$$

Thus in the scalar case we can write down the Milstein's method explicitly as follows.

Algorithm 5.7 (Scalar Milstein's method)**.** *Draw $\hat{x}_0 \sim p(x_0)$ and divide the time interval $[0,t]$ into K steps of length Δt. At each step k do the following:*

1. *Draw random variable $\Delta\beta_k$ from the distribution (where $t_k = k\,\Delta t$)*

$$\Delta\beta_k \sim \mathrm{N}(0, q\,\Delta t). \tag{5.61}$$

2. *Compute*

$$\begin{aligned}\hat{x}(t_{k+1}) = \hat{x}(t_k) &+ f(\hat{x}(t_k), t_k)\,\Delta t + L(\hat{x}(t_k), t_k)\,\Delta\beta_k \\ &+ \frac{1}{2}\frac{\partial L(\hat{x}(t_k), t_k)}{\partial x}\,L(\hat{x}(t_k), t_k)\,(\Delta\beta_k^2 - q\,\Delta t).\end{aligned} \tag{5.62}$$

We could now form even higher order Itô–Taylor series expansions by including more terms into the series. However, if we try to derive higher order methods than Milstein's method, we encounter higher order iterated Itô integrals which will turn out to be very difficult to compute. Fortunately, the additive noise case is much easier and often useful as well.

Let's now consider the case that $\mathbf{L}$ is in fact constant, which implies that $\mathcal{L}_t\mathbf{L} = \mathcal{L}_{\beta,v}\mathbf{L} = 0$. Let's also assume that $\mathbf{Q}$ is constant. In that case Equation (5.48) gives

$$\begin{aligned}\mathbf{x}(t) = \mathbf{x}(t_0) &+ \mathbf{f}(\mathbf{x}(t_0), t_0)\,(t-t_0) + \mathbf{L}(\mathbf{x}(t_0), t_0)\,(\boldsymbol{\beta}(t) - \boldsymbol{\beta}(t_0)) \\ &+ \int_{t_0}^{t}\int_{t_0}^{\tau} \mathcal{L}_t\mathbf{f}(\mathbf{x}(\tau), \tau)\,\mathrm{d}\tau\,\mathrm{d}\tau + \sum_v \int_{t_0}^{t}\int_{t_0}^{\tau} \mathcal{L}_{\beta,v}\mathbf{f}(\mathbf{x}(t), t)\,\mathrm{d}\beta_v\,\mathrm{d}\tau.\end{aligned} \tag{5.63}$$

As the identities in Equation (5.47) are completely general, we can also apply them to $\mathcal{L}_t\mathbf{f}(\mathbf{x}(t), t)$ and $\mathcal{L}_{\beta,v}\mathbf{f}(\mathbf{x}(t), t)$ which gives

$$\begin{aligned}\mathcal{L}_t\mathbf{f}(\mathbf{x}(t), t) &= \mathcal{L}_t\mathbf{f}(\mathbf{x}(t_0), t_0) + \int_{t_0}^{t} \mathcal{L}_t^2\mathbf{f}(\mathbf{x}(t), t)\,\mathrm{d}t \\ &\quad + \sum_v \int_{t_0}^{t} \mathcal{L}_{\beta,v}\mathcal{L}_t\mathbf{f}(\mathbf{x}(t), t)\,\mathrm{d}\beta_v, \\ \mathcal{L}_{\beta,v}\mathbf{f}(\mathbf{x}(t), t) &= \mathcal{L}_{\beta,v}\mathbf{f}(\mathbf{x}(t_0), t_0) + \int_{t_0}^{t} \mathcal{L}_t\mathcal{L}_{\beta,v}\mathbf{f}(\mathbf{x}(t), t)\,\mathrm{d}t \\ &\quad + \sum_v \int_{t_0}^{t} \mathcal{L}_{\beta,v}^2\mathbf{f}(\mathbf{x}(t), t)\,\mathrm{d}\beta_v.\end{aligned} \tag{5.64}$$

By substituting these identities into Equation (5.63) gives

$$\begin{aligned}\mathbf{x}(t) = \mathbf{x}(t_0) &+ \mathbf{f}(\mathbf{x}(t_0), t_0)\,(t - t_0) + \mathbf{L}(\mathbf{x}(t_0), t_0)\,(\boldsymbol{\beta}(t) - \boldsymbol{\beta}(t_0)) \\ &+ \mathcal{L}_t \mathbf{f}(\mathbf{x}(t_0), t_0)\,\frac{(t - t_0)^2}{2} \\ &+ \sum_v \mathcal{L}_{\beta,v} \mathbf{f}(\mathbf{x}(t_0), t_0) \int_{t_0}^{t} [\beta_v(\tau) - \beta_v(t_0)]\,\mathrm{d}\tau + \text{remainder}.\end{aligned} \tag{5.65}$$

Thus the resulting approximation is

$$\begin{aligned}\mathbf{x}(t) \approx \mathbf{x}(t_0) &+ \mathbf{f}(\mathbf{x}(t_0), t_0)\,(t - t_0) + \mathbf{L}(\mathbf{x}(t_0), t_0)\,(\boldsymbol{\beta}(t) - \boldsymbol{\beta}(t_0)) \\ &+ \mathcal{L}_t \mathbf{f}(\mathbf{x}(t_0), t_0)\,\frac{(t - t_0)^2}{2} + \sum_v \mathcal{L}_{\beta,v} \mathbf{f}(\mathbf{x}(t_0), t_0) \int_{t_0}^{t} [\beta_v(\tau) - \beta_v(t_0)]\,\mathrm{d}\tau.\end{aligned} \tag{5.66}$$

Note that the term $\boldsymbol{\beta}(t) - \boldsymbol{\beta}(t_0)$ and the integral $\int_{t_0}^{t} [\beta_v(\tau) - \beta_v(t_0)]\,\mathrm{d}\tau$ really refer to the *same* Brownian motion and thus the terms are correlated. Fortunately in this case both the terms are Gaussian and it is easy to compute their joint distribution:

$$\begin{pmatrix} \int_{t_0}^{t} [\boldsymbol{\beta}(\tau) - \boldsymbol{\beta}(t_0)] \\ \boldsymbol{\beta}(t) - \boldsymbol{\beta}(t_0) \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{Q}\,(t - t_0)^3/3 & \mathbf{Q}\,(t - t_0)^2/2 \\ \mathbf{Q}\,(t - t_0)^2/2 & \mathbf{Q}\,(t - t_0) \end{pmatrix} \right) \tag{5.67}$$

By neglecting the remainder term, we get a strong order 1.5 Itô–Taylor expansion method, which has also been recently studied in the context of filtering theory (Arasaratnam et al., 2010; Särkkä and Solin, 2012).

Algorithm 5.8 (Strong order 1.5 Itô–Taylor method for constant diffusion)**.** *When* $\mathbf{L}$ *and* $\mathbf{Q}$ *are constant, we get the following algorithm. Draw* $\hat{\mathbf{x}}_0 \sim p(\mathbf{x}_0)$ *and divide time interval* $[0, t]$ *into* K *steps of length* Δt*. At each step* k *do the following:*

1. *Draw random variables* $\Delta\boldsymbol{\zeta}_k$ *and* $\Delta\boldsymbol{\beta}_k$ *from the joint distribution*

$$\begin{pmatrix} \Delta\boldsymbol{\zeta}_k \\ \Delta\boldsymbol{\beta}_k \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{Q}\,\Delta t^3/3 & \mathbf{Q}\,\Delta t^2/2 \\ \mathbf{Q}\,\Delta t^2/2 & \mathbf{Q}\,\Delta t \end{pmatrix} \right). \tag{5.68}$$

2. *Compute*

$$\begin{aligned}\hat{\mathbf{x}}(t_{k+1}) = \hat{\mathbf{x}}(t_k) &+ \mathbf{f}(\hat{\mathbf{x}}(t_k), t_k)\,\Delta t + \mathbf{L}\,\Delta\boldsymbol{\beta}_k \\ &+ \mathbf{a}_k\,\frac{(t - t_0)^2}{2} + \sum_v \mathbf{b}_{v,k}\,\Delta\boldsymbol{\zeta}_k,\end{aligned} \tag{5.69}$$

where

$$\begin{aligned}\mathbf{a}_k &= \frac{\partial \mathbf{f}(\hat{\mathbf{x}}(t_k), t_k)}{\partial t} + \sum_u \frac{\partial \mathbf{f}(\hat{\mathbf{x}}(t_k), t_k)}{\partial x_u}\, f_u(\hat{\mathbf{x}}(t_k), t_k) \\ &\quad + \frac{1}{2} \sum_{uv} \frac{\partial^2 \mathbf{f}(\hat{\mathbf{x}}(t_k), t_k)}{\partial x_u\, \partial x_v}\, [\mathbf{L}\,\mathbf{Q}\,\mathbf{L}^\mathsf{T}]_{uv} \\ \mathbf{b}_{v,k} &= \sum_u \frac{\partial \mathbf{f}(\hat{\mathbf{x}}(t_k), t_k)}{\partial x_u}\, \mathbf{L}_{uv}.\end{aligned} \tag{5.70}$$

As an interesting note on higher order iterated Itô integrals, we point out the recursive relation originally published by Itô himself. Applying the Itô formula for n times iterated Itô integrals, leads to the following recursion:

$$n! \int_{t_0}^{t} \int_{t_0}^{\tau_n} \dots \int_{t_0}^{\tau_2} \mathrm{d}\beta(\tau_1)\, \mathrm{d}\beta(\tau_2) \dots \mathrm{d}\beta(\tau_n) = (t - t_0)^{n/2}\, \mathrm{H}_n\left(\frac{\beta(t) - \beta(t_0)}{\sqrt{t - t_0}}\right), \tag{5.71}$$

where $\mathrm{H}_n(t)$ denotes the probabilists' Hermite polynomials ($\mathrm{H}_0(t) = 1, \mathrm{H}_1(t) = t, \mathrm{H}_2(t) = t^2 - 1, \mathrm{H}_3(t) = t^3 - 3t, \dots$). They are defined through the recursion $\mathrm{H}_{n+1}(t) = t\,\mathrm{H}_n(t) - \mathrm{d}/\mathrm{d}t\ \mathrm{H}_n(t)$. The result in Equation (5.60) can easily be verified from the formula above.

5.5 Weak approximations of Itô–Taylor series

The interest in solving the SDE is not always in the solution trajectories. Usually more interest is put into the distribution of the trajectories at a given time point rather than their paths. Thus we might be interested in forming approximations that describe accurately enough the probability distribution of the trajectories. Weak approximations of the Itô process, that is a process with approximately the same probability distribution, provide much more freedom in forming the approximations.

For instance, we can replace the initial value $\mathbf{x}_0$ with some appropriate probability distribution, or more importantly we can replace the random increments $\Delta\boldsymbol{\beta}_k$ with more convenient approximations $\Delta\hat{\boldsymbol{\beta}}_k$ with similar moment properties.

The kind of approximations required here are much weaker than those required by the strong convergence criterion. The *weak order of convergence* can be defined to be the smallest exponent α such that

$$|\,\mathrm{E}\,[g(\mathbf{x}(t_n))] - \mathrm{E}\,[g(\hat{\mathbf{x}}(t_n))]\,| \leq K\,\Delta t^{\alpha}, \tag{5.72}$$

for any polynomial function g. When the diffusion coefficient vanishes ($\mathbf{L}(\mathbf{x}, t) \equiv 0$), this weak convergence criterion with $g(\mathbf{x}) = \mathbf{x}$ reduces to the usual deterministic convergence criterion for ordinary differential equations—as does the criterion for strong convergence.

For weak convergence, we only need to approximate the measure induced by the Itô process $\mathbf{x}(t)$, so we can replace the Gaussian increments by other random variables with similar moment properties. Considering this, we can replace the increments in Algorithm 5.5. This leads to the *simplified weak Euler–Maruyama scheme*

$$\hat{\mathbf{x}}(t_{k+1}) = \hat{\mathbf{x}}(t_k) + \mathbf{f}\,(\hat{\mathbf{x}}(t_k), t_k)\,\Delta t + \mathbf{L}(\hat{\mathbf{x}}(t_k), t_k)\,\Delta\hat{\boldsymbol{\beta}}_k, \tag{5.73}$$

where the $\Delta\hat{\beta}_k^j$, $j = 1, 2, \dots, m$, must be independent random variables fulfilling suitable moment conditions. For example, we could use the following two-point distributed random variables (Kloeden and Platen, 1999)

$$\mathrm{P}(\Delta\hat{\beta}_k^j = \pm\sqrt{\Delta t}) = \frac{1}{2}. \tag{5.74}$$

As discussed earlier, the Euler–Maruyama scheme has strong convergence order $\gamma = 0.5$, but weak order $\alpha = 1.0$ (provided certain differentiability conditions are met).

As we noticed in the previous section, multi-dimensional and higher order Itô–Taylor approximations also involve additional random variables and iterated Itô integrals, which make them difficult in practice. The same applies to weak Itô–Taylor approximations, but handling these is much simpler than in the case of strong approximations. As a rule of thumb, Kloeden and Platen (1999) state that an Itô–Taylor approximation converges with any desired weak order $\alpha = 1.0, 2.0, \ldots$, when the number of stochastic integrals up to multiplicity α are included in the expansion. As an example they give the following scalar time-invariant weak order $\alpha = 2.0$ scheme:

Algorithm 5.9 (Scalar weak order 2.0 Itô–Taylor method)**.** *Draw $\hat{x}_0 \sim p(x_0)$ and divide the time interval $[0, t]$ into K steps of length Δt. At each step k do the following:*

$$\begin{aligned}\hat{x}(t_{k+1}) = \hat{x}(t_k) &+ f(\hat{x}(t_k))\,\Delta t + L(\hat{x}(t_k))\,\Delta\hat{\beta}_k \\ &+ \frac{1}{2}L(\hat{x}(t_k))\,\frac{\partial L(\hat{x}(t_k))}{\partial x}((\Delta\hat{\beta}_k)^2 - \Delta t) \\ &+ \frac{\partial f(\hat{x}(t_k))}{\partial x}\,L(\hat{x}(t_k))\,\Delta\hat{\zeta}_k \\ &+ \frac{1}{2}\left(f(\hat{x}(t_k))\,\frac{\partial f(\hat{x}(t_k))}{\partial x} + \frac{1}{2}\frac{\partial^2 f(\hat{x}(t_k))}{\partial x^2}\,L^2(\hat{x}(t_k))\right)(\Delta t)^2 \\ &+ \left(f(\hat{x}(t_k))\,\frac{\partial L(\hat{x}(t_k))}{\partial x} + \frac{1}{2}\frac{\partial^2 L(\hat{x}(t_k))}{\partial x^2}\,L^2(\hat{x}(t_k))\right)\left(\Delta\hat{\beta}_k\,\Delta t - \Delta\hat{\zeta}_k\right).\end{aligned} \tag{5.75}$$

Here $\Delta\hat{\beta}_k$ and $\Delta\hat{\zeta}_k$ approximates $\Delta\beta_k$ and $\Delta\zeta_k$. We can choose

$$\Delta\hat{\beta}_k = \Delta\beta_k \quad \textit{and} \quad \Delta\hat{\zeta}_k = \frac{1}{2}\Delta\beta_k\,\Delta t \tag{5.76}$$

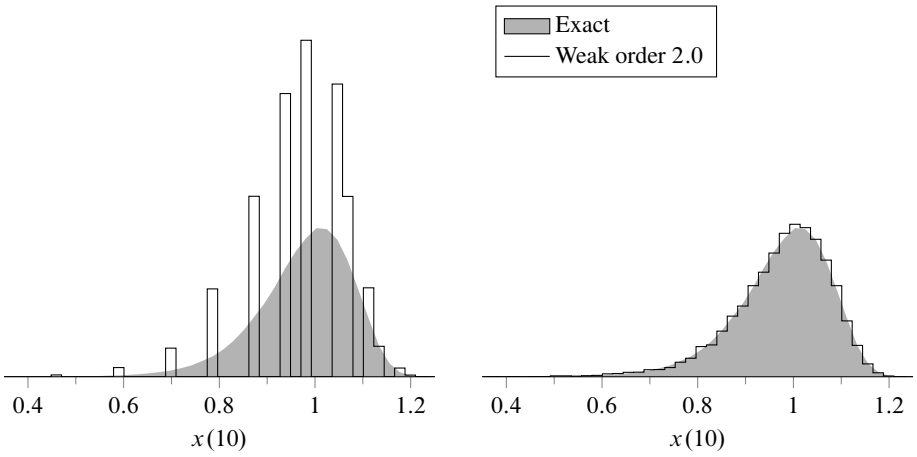
with $\Delta\beta_k \sim \mathrm{N}(0, \Delta t)$, or instead of considering the normal increments, we could use

$$\Delta\hat{\beta}_k = (\Delta t)^{1/2}\,\theta_k \quad \textit{and} \quad \Delta\hat{\zeta}_k = \frac{1}{2}(\Delta t)^{3/2}\,\theta_k, \tag{5.77}$$

where θ_ks are independent three-point distributed random variables with

$$\mathrm{P}(\theta_k = \pm\sqrt{3}) = \frac{1}{6} \quad \textit{and} \quad \mathrm{P}(\theta_k = 0) = \frac{2}{3}. \tag{5.78}$$

As an example of an application of the above algorithm we provide the following.



(a) Three-point increments **(b)** Gaussian increments

Figure 5.2: The exact distribution at $t = 10$, and histograms of the 10,000 samples simulated by the weak order 2.0 scheme both using the three-point distributed and Gaussian increments. Both solutions **(a)** and **(b)** have similar moment properties.

Example 5.2. *We return to the non-linear SDE model considered earlier in this chapter in Example 5.1. The first and second derivatives of the drift and diffusion functions are:*

$$\frac{\mathrm{d}f(x)}{\mathrm{d}x} = -\frac{1}{100}\cos^2(x)\,(2\cos(2x) - 1), \qquad \frac{\mathrm{d}L(x)}{\mathrm{d}x} = -\frac{1}{5}\sin(x)\,\cos(x),$$
$$\frac{\mathrm{d}^2 f(x)}{\mathrm{d}x^2} = \frac{1}{100}(\sin(2x) + 2\sin(4x)), \qquad \frac{\mathrm{d}^2 L(x)}{\mathrm{d}x^2} = -\frac{1}{5}\cos(2x).$$

We apply the weak order 2.0 scheme in Algorithm 5.9 to this problem, and characterize the solution at $t = 10$. We use a large step size $\Delta t = 1$ and simulate 10,000 trajectories. Figure 5.2 shows histograms of the values at $\hat{x}(10)$ both using the three-point distributed increments and Gaussian increments. Even though both solutions have similar moment properties (mean, variance, skewness, kurtosis, ...), the Gaussian increments appear nicer in the visualization.

Chapter 6

Stochastic Runge–Kutta methods

This chapter is concerned with solving arbitrary stochastic differential equations by numerical approximations. We will focus on *Runge–Kutta methods* which are an important family of explicit and implicit iterative methods for numerical approximative solving of differential equations. Their applicability is mostly motivated by their plug-and-play formulations, which are typically derivative-free, and only requires specification of the (stochastic) differential equation.

The field of solving ordinary differential equations by numerical approximations has been extensively studied during the past century, and for practical use there exist highly optimized implementations in all major software packages. However, methods for solving stochastic differential equations are harder to find and not too many implementations are readily available.

Even though we focus on Runge–Kutta methods, it is worth mentioning that not all numerical methods fall under their formulation. Some other methods that do not fit under the framework of Runge–Kutta schemes are (i) *multistep methods* where the information form intermediate steps is not discarded, but rather re-used (*e.g.*, the family of *Adams methods*); (ii) *multiderivative methods*, that is, methods that use the derivatives (gradient) of the integrand in addition to the integrand itself; (iii) *higher-order methods*, that is, methods where conversion of a higher-order problem to a first-order differential equation is not desired (*e.g.*, the *Nyström method* fits well with second-order ODEs); (iv) *tailored methods* that use some special properties of the particular differential equation. We will not discuss these methods in detail, but the interested reader can read about them, for example, in the book by Hairer et al. (2008). Many of these methods have some sort of stochastic equivalent, but those are not discussed in this material either.

To ease the formulation of stochastic Runge–Kutta methods, we will first go through some background on ordinary Runge–Kutta methods. After that we focus on strong stochastic methods, and thereafter consider some weak methods. We will try to provide general tools, but it is worth noting that often in the case of stochastic differential equations special structure of the problem can have a large impact on the complexity of the solution method.

6.1 Runge–Kutta methods for ODEs

Runge–Kutta (RK) methods are an important family of iterative methods for the approximation of solutions of ordinary differential equations. The name stems from the German mathematicians Carl Runge and Martin Wilhelm Kutta, whose work many of the modern-day methods build upon.

The simplest Runge–Kutta method we can think of is the (forward) Euler scheme (see Alg. 1.1) which is based on sequential linearization of the ODE system. This method is easy to understand and implement, but the global error of the method depends linearly on the step size Δt. The innovation Runge came up with was that subdivision of the integration interval into intermediate steps (as had earlier been done in quadrature methods, *e.g.* the *midpoint rule*, where the integrand was independent of t), can help build more efficient methods. Such higher order methods can reach the same precision with fewer steps, which makes them appealing.

Consider the first-order non-linear ODE from Chapter 1

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}(t), t), \qquad \mathbf{x}(t_0) = \text{given}, \tag{6.1}$$

which can be integrated to give

$$\mathbf{x}(t) = \mathbf{x}(t_0) + \int_{t_0}^{t} \mathbf{f}(\mathbf{x}(\tau), \tau)\, \mathrm{d}\tau. \tag{6.2}$$

Recall from the previous chapter that we used a *Taylor series expansion* for the solution of the ODE

$$\begin{aligned} \mathbf{x}(t) = \mathbf{x}(t_0) &+ \mathbf{f}(\mathbf{x}(t_0), t_0)\,(t - t_0) \\ &+ \frac{1}{2!}\,\mathscr{L}\,\mathbf{f}(\mathbf{x}(t_0), t_0)\,(t - t_0)^2 \\ &+ \frac{1}{3!}\,\mathscr{L}^2\,\mathbf{f}(\mathbf{x}(t_0), t_0)\,(t - t_0)^3 \\ &+ \cdots, \end{aligned} \tag{6.3}$$

where we used the linear operator

$$\mathscr{L}(\bullet) = \frac{\partial}{\partial t}(\bullet) + \sum_i f_i\, \frac{\partial}{\partial x_i}(\bullet) \tag{6.4}$$

to come up with the convenient formulation of the series. Thus the series expansion

is equal to

$$\begin{aligned}\mathbf{x}(t) = \mathbf{x}(t_0) &+ \mathbf{f}(\mathbf{x}(t_0), t_0)\,(t - t_0) \\ &+ \frac{1}{2!}\left\{\frac{\partial}{\partial t}\mathbf{f}(\mathbf{x}(t_0), t_0) + \sum_i f_i(\mathbf{x}(t_0), t_0)\,\frac{\partial}{\partial x_i}\mathbf{f}(\mathbf{x}(t_0), t_0)\right\}(t - t_0)^2 \\ &+ \frac{1}{3!}\left\{\frac{\partial[\mathscr{L}\,\mathbf{f}(\mathbf{x}(t_0), t_0)]}{\partial t} + \sum_i f_i(\mathbf{x}(t_0), t_0)\,\frac{\partial[\mathscr{L}\,\mathbf{f}(\mathbf{x}(t_0), t_0)]}{\partial x_i}\right\}(t - t_0)^3 \\ &+ \cdots \end{aligned} \tag{6.5}$$

If we were only to consider the terms up to Δt, we would recover the Euler method, which clearly is a derivative-free Runge–Kutta scheme. However, here we wish to get hold of higher-order methods. For the sake of simplicity we now stop at the term $(t - t_0)^2 = (\Delta t)^2$, and write

$$\begin{aligned}\mathbf{x}(t_0 + \Delta t) \approx \mathbf{x}(t_0) &+ \mathbf{f}(\mathbf{x}(t_0), t_0)\,\Delta t \\ &+ \frac{1}{2}\left\{\frac{\partial}{\partial t}\mathbf{f}(\mathbf{x}(t_0), t_0) + \sum_i f_i(\mathbf{x}(t_0), t_0)\,\frac{\partial}{\partial x_i}\mathbf{f}(\mathbf{x}(t_0), t_0)\right\}(\Delta t)^2.\end{aligned} \tag{6.6}$$

This equation still contains derivatives, and we aim to get rid of them and be able to write the expression in terms of the function $\mathbf{f}(\cdot, \cdot)$ evaluated at various points. That is, we now seek a form:

$$\begin{aligned}\mathbf{x}(t_0 + \Delta t) \approx \mathbf{x}(t_0) &+ A\,\mathbf{f}(\mathbf{x}(t_0), t_0)\,\Delta t \\ &+ B\,\mathbf{f}(\mathbf{x}(t_0) + C\,\mathbf{f}(\mathbf{x}(t_0), t_0)\,\Delta t, t_0 + D\,\Delta t)\,\Delta t,\end{aligned} \tag{6.7}$$

where A, B, C, and D are unknown. In the last term, we can instead consider the truncated Taylor expansion (linearization) around $\mathbf{f}(\mathbf{x}(t_0), t_0)$ with the chosen increments as follows:

$$\begin{aligned}&\mathbf{f}(\mathbf{x}(t_0) + C\,\mathbf{f}(\mathbf{x}(t_0), t_0)\,\Delta t, t_0 + D\,\Delta t) = \mathbf{f}(\mathbf{x}(t_0), t_0) \\ &+ C\left(\sum_i f_i(\mathbf{x}(t_0), t_0)\,\frac{\partial}{\partial x_i}\mathbf{f}(\mathbf{x}(t_0), t_0)\right)\Delta t + D\,\frac{\partial \mathbf{f}(\mathbf{x}(t_0), t_0)}{\partial t}\,\Delta t + \cdots\end{aligned} \tag{6.8}$$

Combining the above two equations gives:

$$\begin{aligned}\mathbf{x}(t_0 + \Delta t) \approx \mathbf{x}(t_0) &+ (A + B)\,\mathbf{f}(\mathbf{x}(t_0), t_0)\,\Delta t \\ &+ B\left\{C\sum_i f_i(\mathbf{x}(t_0), t_0)\,\frac{\partial}{\partial x_i}\mathbf{f}(\mathbf{x}(t_0), t_0) + D\,\frac{\partial \mathbf{f}(\mathbf{x}(t_0), t_0)}{\partial t}\right\}(\Delta t)^2.\end{aligned} \tag{6.9}$$

If we now compare the above equation to the original truncated Taylor expansion in Equation (6.6), we get the following conditions for our coefficients:

$$A + B = 1, \quad B = \frac{1}{2}, \quad C = 1, \quad \text{and} \quad D = 1. \tag{6.10}$$

We can now solve $A = 1/2$. Thus the approximative step given by this method can be written as

$$\hat{\mathbf{x}}(t_0 + \Delta t) = \mathbf{x}(t_0) + \frac{1}{2}\big[\mathbf{f}(\tilde{\mathbf{x}}_1, t_0) + \mathbf{f}(\tilde{\mathbf{x}}_2, t_0 + \Delta t)\big]\,\Delta t, \tag{6.11}$$

where the supporting values are given by

$$\begin{aligned} \tilde{\mathbf{x}}_1 &= \mathbf{x}(t_0), \\ \tilde{\mathbf{x}}_2 &= \mathbf{x}(t_0) + \mathbf{f}(\tilde{\mathbf{x}}_1, t_0)\,\Delta t. \end{aligned} \tag{6.12}$$

What we derived here is a two-stage method (actually the Heun's method presented in Alg. 1.2) with the finite differences determined by the choices we did in truncating the series expansion. The choices of how and what to truncate determine the number of terms in the expansion, and thus also affects the number of equations to solve. Coming up with higher-order methods becomes increasingly complicated with the number of terms. The general principle however remains the same, and Runge–Kutta methods are constructed by evaluating the model function (often recursively) at a discrete number of step sizes, and weighting these evaluations.

We write down a general s-stage algorithm for Runge–Kutta methods:

Algorithm 6.1 (Runge–Kutta methods)**.** *Start from* $\hat{\mathbf{x}}(t_0) = \mathbf{x}(t_0)$ *and divide the integration interval* $[t_0, t]$ *into* n *steps* $t_0 < t_1 < t_2 < \ldots < t_n = t$ *such that* $\Delta t = t_{k+1} - t_k$*. The integration method is defined by its* Butcher tableau*:*

$$\begin{array}{c|c} \mathbf{c} & \mathbf{A} \\ \hline & \boldsymbol{\alpha}^{\mathsf{T}} \end{array} \tag{6.13}$$

On each step k *approximate the solution as follows:*

$$\hat{\mathbf{x}}(t_{k+1}) = \hat{\mathbf{x}}(t_k) + \sum_{i=1}^{s} \alpha_i\, \mathbf{f}(\tilde{\mathbf{x}}_i, \tilde{t}_i)\,\Delta t, \tag{6.14}$$

where $\tilde{t}_i = t_k + c_i \Delta t$ *and* $\tilde{\mathbf{x}}_i = \hat{\mathbf{x}}(t_k) + \sum_{j=1}^{s} A_{i,j}\, \mathbf{f}(\tilde{\mathbf{x}}_j, \tilde{t}_j)\,\Delta t$.

As can be interpreted from above, ordinary Runge–Kutta methods are commonly expressed in terms of a table called the *Butcher tableau*:

$$\begin{array}{c|cccc} c_1 & A_{1,1} & & & \\ c_2 & A_{2,1} & A_{2,2} & & \\ \vdots & \vdots & & \ddots & \\ c_s & A_{s,1} & A_{s,2} & \ldots & A_{s,s} \\ \hline & \alpha_1 & \alpha_2 & \ldots & \alpha_s \end{array} \tag{6.15}$$

An explicit Runge–Kutta method is said to be consistent if $\sum_{j=1}^{i-1} A_{i,j} = c_i$, for $i = 2, 3, \ldots, s$.

We present the Butcher tableau for two common Runge–Kutta methods.. The first method is the forward Euler method (see Alg. 1.1) and the second the well-known fourth order classical Runge–Kutta method (see Alg 1.3):

Example 6.1 (Forward Euler)**.** *The forward Euler scheme in Algorithm 1.1 has the Butcher tableau:*

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array} \tag{6.16}$$

which gives the recursion $\hat{\mathbf{x}}(t_{k+1}) = \hat{\mathbf{x}}(t_k) + \mathbf{f}(\hat{\mathbf{x}}(t_k), t_k)\,\Delta t$.

Example 6.2 (The fourth-order Runge–Kutta method)**.** *The well-known RK4 method in Algorithm 1.3 has the following Butcher tableau:*

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array} \tag{6.17}$$

The Runge–Kutta schemes considered above, are all so-called explicit schemes. Explicit schemes can suffer from numerical instability, when the solution includes rapidly varying terms. Such problems are called *stiff* equations. Stiff equations require explicit schemes to use small step sizes in order to not diverge from a solution path.

A better suited family for stiff problems are the so-called *implicit* Runge–Kutta methods, which provide additional stability to the iterative solution. For implicit methods, the Buther tableau is no longer lower-triangular, but the tableau can be full. The consequence of using a full table is that at every step, a system of algebraic equations has to be solved. This increases the computational cost considerably. The advantage of implicit Runge–Kutta methods over explicit ones is their greater stability, especially when applied to stiff equations.

The simplest example of an implicit method is the *backward Euler* scheme:

Example 6.3 (Backward Euler)**.** *The implicit backward Euler scheme has the Butcher tableau:*

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array} \tag{6.18}$$

which gives the recursion $\hat{\mathbf{x}}(t_{k+1}) = \hat{\mathbf{x}}(t_k) + \mathbf{f}(\hat{\mathbf{x}}(t_{k+1}), t_k + \Delta t)\,\Delta t$.

There are a lot of further topics to consider in Runge–Kutta methods, such as stability analysis and adaptive step size methods. We, however, will not discuss these, but try to move to the stochastic Runge–Kutta methods.

Example 6.4. *We study the two-dimensional non-linear ordinary differential equation system*

$$\begin{aligned} \dot{x}_1 &= x_1 - x_2 - x_1^3, \\ \dot{x}_2 &= x_1 + x_2 - x_2^3, \end{aligned} \tag{6.19}$$

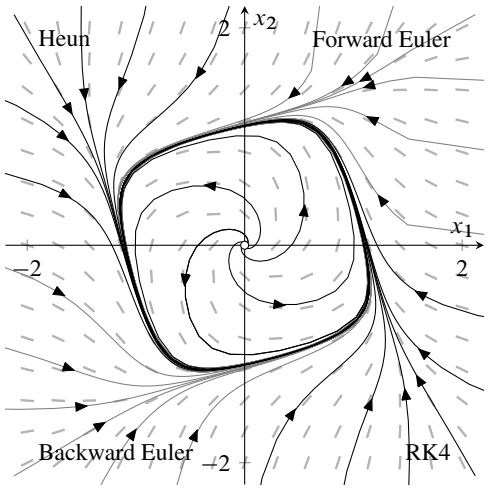


Figure 6.1: Demonstration of four Runge–Kutta schemes for the same problem with a step size of $\Delta t = 2^{-4}$.

This system has only one fixed point, $\mathbf{x} = (0, 0)$ *(an unstable spiral), but it also has a limit cycle. We aim to test various Runge–Kutta methods by simulating trajectories of this problem.*

We use a time-span of $[0, 10]$*, with a step size of* $\Delta t = 2^{-4}$*. The methods demonstrated are the forward Euler method, Heun's method, the backward (implicit) Euler method, and the classical fourth-order Runge–Kutta method. Figure 6.1 shows six trajectories for each method, starting from respective quadrants. The results should be symmetrical, but especially the forward Euler results do not match the rest.*

6.2 Strong stochastic Runge–Kutta methods

A practical disadvantage of the Taylor approximations considered in the previous chapter is that the derivatives of various orders of the drift and diffusion functions must be determined and evaluated at each step. However, there are discrete time approximations which avoid the use of derivatives. They are in general referred to as *stochastic Runge–Kutta methods*. Stochastic versions of the Runge–Kutta methods are not as simple as in the case of deterministic equations.

As has been discussed earlier, the *Euler–Maryama* scheme can easily be constructed by discretizing the time interval and formulating the SDE as a recursive algorithm. Thus the Euler–Maruyama scheme can be seen as the simplest stochastic Runge–Kutta method—similarly as we interpreted the Euler method as a simple ordinary Runge–Kutta scheme.

In practice, a higher-order stochastic Runge–Kutta method can be derived, for example, by replacing the closed-form derivatives in the Milstein's method

(Algs. 5.6 or 5.7) with suitable finite differences (see Kloeden et al., 1994; Kloeden and Platen, 1999). So, if we heuristically replace the partial differential in Algorithm 5.7 with a finite difference, we can write a method for scalar time-invariant models:

$$\hat{x}(t_{k+1}) = \hat{x}(t_k) + f(x(t_k))\,\Delta t + L(\hat{x}_k)\,\Delta\beta_k + \frac{1}{2\sqrt{\Delta t}}\big[L(\tilde{x}) - L(\hat{x})\big]\big[(\Delta\beta_k)^2 - \Delta t\big] \quad (6.20)$$

with supporting value $\tilde{x} = \hat{x}_k + L(\hat{x}_k)\sqrt{\Delta t}$. where we consider only standard Brownian motions. This method is of strong order 1.0.

However, we still *cannot get rid of the iterated Itô integral* occurring in Milstein's method. An important thing to note is that stochastic versions of Runge–Kutta methods cannot be derived as simple extensions of the deterministic Runge–Kutta methods—see Burrage et al. (2006) which is a response to the article by Wilkie (2004).

To provide a more widely applicable perspective on the methods, we follow a similar derivation as we did for the ordinary Runge–Kutta methods. Recall the following multi-dimensional SDE formulation

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}(t), t)\,\mathrm{d}t + \mathbf{L}(\mathbf{x}(t), t)\,\mathrm{d}\boldsymbol{\beta}, \qquad \mathbf{x}(t_0) \sim p(\mathbf{x}(t_0)), \quad (6.21)$$

where the drift is defined by $\mathbf{f} : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ and the diffusion coefficients by $\mathbf{L} : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d \times \mathbb{R}^m$, and the driving noise process $\boldsymbol{\beta}(t) = (\beta^{(1)}(t), \beta^{(2)}(t), \ldots, \beta^{(m)}(t))$ is an m-dimensional standard Brownian motion. In integral form the equation can be expressed as

$$\mathbf{x}(t) = \mathbf{x}(t_0) + \int_{t_0}^{t} \mathbf{f}(\mathbf{x}(\tau), \tau)\,\mathrm{d}\tau + \int_{t_0}^{t} \mathbf{L}(\mathbf{x}(\tau), \tau)\,\mathrm{d}\boldsymbol{\beta}(\tau). \quad (6.22)$$

As we saw in the previous chapter, applying the Itô formula to the terms $\mathbf{f}(\mathbf{x}(t), t)$ and $\mathbf{L}(\mathbf{x}(t), t)$ and collecting the terms gives an Itô–Taylor series expansion of the solution:

$$\begin{aligned}\mathbf{x}(t) = \mathbf{x}(t_0) &+ \mathbf{f}(\mathbf{x}(t_0), t_0)\,(t - t_0) + \mathbf{L}(\mathbf{x}(t_0), t_0)\,(\boldsymbol{\beta}(t) - \boldsymbol{\beta}(t_0)) \\ &+ \int_{t_0}^{t}\int_{t_0}^{\tau} \mathcal{L}_t \mathbf{f}(\mathbf{x}(\tau), \tau)\,\mathrm{d}\tau\,\mathrm{d}\tau \\ &+ \sum_i \int_{t_0}^{t}\int_{t_0}^{\tau} \mathcal{L}_{\beta,i} \mathbf{f}(\mathbf{x}(\tau), \tau)\;\mathrm{d}\beta^{(i)}(\tau)\,\mathrm{d}\tau \\ &+ \int_{t_0}^{t}\int_{t_0}^{\tau} \mathcal{L}_t \mathbf{L}(\mathbf{x}(\tau), \tau)\;\mathrm{d}\tau\,\mathrm{d}\boldsymbol{\beta}(\tau) \\ &+ \sum_i \int_{t_0}^{t}\int_{t_0}^{\tau} \mathcal{L}_{\beta,i} \mathbf{L}(\mathbf{x}(\tau), \tau)\;\mathrm{d}\beta^{(i)}(\tau)\,\mathrm{d}\boldsymbol{\beta}(\tau). \end{aligned} \quad (6.23)$$

Similarly as we did in the previous section, we can consider truncated series expansions of various degrees for each of these terms. Collecting the terms can give a similar kind of formulation in terms of tabulated values as we did for the ordinary RK schemes. The extra terms involving the iterated and cross-term Itô integrals complicate the formulation.

Rößler (2010) considers a general class of multi-dimensional strong order 1.0 stochastic Runge–Kutta schemes, where iterated integrals are avoided in the scheme and they only appear in the supporting values. A more general formulation is given in the next section, where weak order methods are considered. The general multi-dimensional schemes by Rößler are as follows:

Algorithm 6.2 (A class of stochastic Runge–Kutta methods of strong order 1.0)**.** *Start from* $\hat{\mathbf{x}}(t_0) = \mathbf{x}(t_0)$ *and divide the integration interval* $[t_0, t]$ *into* n *steps* $t_0 < t_1 < t_2 < \ldots < t_n = t$ *such that* $\Delta t = t_{k+1} - t_k$*. The integration method is characterized by its* extended Butcher tableau*:*

$$\begin{array}{c|c|c|c} \mathbf{c}^{(0)} & \mathbf{A}^{(0)} & \mathbf{B}^{(0)} & \\ \hline \mathbf{c}^{(1)} & \mathbf{A}^{(1)} & \mathbf{B}^{(1)} & \\ \hline & \boldsymbol{\alpha}^{\mathsf{T}} & \left[\boldsymbol{\gamma}^{(1)}\right]^{\mathsf{T}} & \left[\boldsymbol{\gamma}^{(2)}\right]^{\mathsf{T}} \end{array} \tag{6.24}$$

On each step k *approximate the solution trajectory as follows:*

$$\begin{aligned} \hat{\mathbf{x}}(t_{k+1}) = \hat{\mathbf{x}}(t_k) &+ \sum_{i=1}^{s} \alpha_i \, \mathbf{f}\left(\tilde{\mathbf{x}}_i^{(0)}, t_k + c_i^{(0)} \Delta t\right) \Delta t \\ &+ \sum_{i=1}^{s} \sum_{n=1}^{m} \left(\gamma_i^{(1)} \Delta \beta_k^{(n)} + \gamma_i^{(2)} \sqrt{\Delta t}\right) \mathbf{L}^n\left(\tilde{\mathbf{x}}_i^{(n)}, t_k + c_i^{(1)} \Delta t\right) \end{aligned} \tag{6.25}$$

with the supporting values

$$\begin{aligned} \tilde{\mathbf{x}}_i^{(0)} = \hat{\mathbf{x}}(t_k) &+ \sum_{j=1}^{s} A_{i,j}^{(0)} \, \mathbf{f}\left(\tilde{\mathbf{x}}_j^{(0)}, t_k + c_j^{(0)} \Delta t\right) \Delta t \\ &+ \sum_{j=1}^{s} \sum_{l=1}^{m} B_{i,j}^{(0)} \, \mathbf{L}^l\left(\tilde{\mathbf{x}}_j^{(l)}, t_k + c_j^{(1)} \Delta t\right) \Delta \beta_k^{(l)}, \end{aligned} \tag{6.26}$$

$$\begin{aligned} \tilde{\mathbf{x}}_i^{(n)} = \hat{\mathbf{x}}(t_k) &+ \sum_{j=1}^{s} A_{i,j}^{(1)} \, \mathbf{f}\left(\tilde{\mathbf{x}}_j^{(0)}, t_k + c_j^{(0)} \Delta t\right) \Delta t \\ &+ \sum_{j=1}^{s} \sum_{l=1}^{m} B_{i,j}^{(1)} \, \mathbf{L}^l\left(\tilde{\mathbf{x}}_j^{(l)}, t_k + c_j^{(1)} \Delta t\right) \frac{\Delta \beta_k^{(l,n)}}{\sqrt{\Delta t}}, \end{aligned} \tag{6.27}$$

for $i = 1, 2, \ldots, s$ *and* $n = 1, 2, \ldots, m$.

The increments in the above algorithm are given by the Itô integrals:

$$\Delta\beta_k^{(i)} = \int_{t_k}^{t_{k+1}} \mathrm{d}\beta^{(i)}(\tau) \quad \text{and} \tag{6.28}$$

$$\Delta\beta_k^{(i,j)} = \int_{t_k}^{t_{k+1}} \int_{t_k}^{\tau_2} \mathrm{d}\beta^{(i)}(\tau_1)\,\mathrm{d}\beta^{(j)}(\tau_2), \tag{6.29}$$

for $i, j = 1, 2, \ldots, m$. The increments $\Delta\beta_k^{(i)}$ are independent normally distributed random variables, $\Delta\beta_k^{(i)} \sim \mathrm{N}(0, \Delta t)$. The iterated stochastic Itô integrals $\Delta\beta_k^{(i,j)}$ are trickier. For these methods, when $i = j$, the multiple Itô integrals can be rewritten as

$$\Delta\beta_k^{(i,i)} = \frac{1}{2}\left(\left[\Delta\beta_k^{(i)}\right]^2 - \Delta t\right), \tag{6.30}$$

which follows from the results given in Equation (5.71). This also generalizes to higher orders. Exact simulation of the integrals $\Delta\beta_k^{(i,j)}$, when $i \neq j$, is not possible, but can be approximated. See Wiktorsson (2001) for an approximative scheme, and Gilsing and Shardlow (2007) for implementation details.

Example 6.5 (Euler–Maruyama Butcher tableau)**.** *The Euler–Maruyama method has the extended Butcher tableau:*

$$\begin{array}{c|cc|c} 0 & 0 & 0 & \\ \hline 0 & 0 & 0 & \\ \hline & 1 & 1 & 0 \end{array} \tag{6.31}$$

and as we recall from the previous chapter, it is of strong order 0.5*.*

Coming up with useful and valid stochastic Runge–Kutta schemes is a delicate process, which we will not consider here. Instead we go through a rather efficient and general scheme proposed by Rößler (2010) and which can be formulated as follows:

Algorithm 6.3 (Strong order 1.0 stochastic Runge–Kutta due to Rößler)**.** *Consider a stochastic Runge–Kutta method with the following extended Butcher tableau:*

$$\begin{array}{c|ccc|ccc|ccc} 0 & & & & & & & & & \\ 1 & 1 & & & 0 & & & & & \\ 0 & 0 & 0 & & 0 & 0 & & & & \\ \hline 0 & & & & & & & & & \\ 1 & 1 & & & 1 & & & & & \\ 1 & 1 & 0 & & -1 & 0 & & & & \\ \hline & \frac{1}{2} & \frac{1}{2} & 0 & 1 & 0 & 0 & 0 & \frac{1}{2} & -\frac{1}{2} \end{array} \tag{6.32}$$

which corresponds to the following iterative scheme

$$\begin{aligned}\hat{\mathbf{x}}(t_{k+1}) = \hat{\mathbf{x}}(t_k) &+ \frac{1}{2}\big\{\mathbf{f}\,(\hat{\mathbf{x}}(t_k), t_k) + \mathbf{f}\,(\tilde{\mathbf{x}}_2^{(0)}, t_k + \Delta t)\big\}\,\Delta t \\ &+ \sum_{n=1}^{m} \Big\{\Delta\beta_k^{(n)}\,\mathbf{L}^n(\hat{\mathbf{x}}(t_k), t_k) \\ &\qquad + \frac{1}{2}\sqrt{\Delta t}\,\big(\mathbf{L}^n(\tilde{\mathbf{x}}_2^{(n)}, t_k + \Delta t) - \mathbf{L}^n(\tilde{\mathbf{x}}_3^{(n)}, t_k + \Delta t)\big)\Big\}\end{aligned} \tag{6.33}$$

with the supporting values $\tilde{\mathbf{x}}_2^{(0)} = \hat{\mathbf{x}}(t_k) + \mathbf{f}\,(\hat{\mathbf{x}}(t_k), t_k)\,\Delta t$*, and*

$$\tilde{\mathbf{x}}_2^{(n)} = \hat{\mathbf{x}}(t_k) + \mathbf{f}\,(\hat{\mathbf{x}}(t_k), t_k)\,\Delta t + \sum_{l=1}^{m} \mathbf{L}^l(\hat{\mathbf{x}}(t_k), t_k)\,\frac{\Delta\beta_k^{(l,n)}}{\sqrt{\Delta t}}, \tag{6.34}$$

$$\tilde{\mathbf{x}}_3^{(n)} = \hat{\mathbf{x}}(t_k) + \mathbf{f}\,(\hat{\mathbf{x}}(t_k), t_k)\,\Delta t - \sum_{l=1}^{m} \mathbf{L}^l(\hat{\mathbf{x}}(t_k), t_k)\,\frac{\Delta\beta_k^{(l,n)}}{\sqrt{\Delta t}}. \tag{6.35}$$

Higher-order methods can be formulated by considering more terms in the Itô–Taylor expansion. This, however, might not be very practical, as the number of required function evaluations grows, as does the complexity of the scheme. However, for models with some special structure this might still be feasible. Examples of such cases are models with commutative noise, additive noise models, where $\mathbf{L}(\mathbf{x}, t) \equiv \mathbf{L}(t)$, or diagonal noise models.

A number of stochastic Runge–Kutta methods have also been presented by Kloeden et al. (1994); Kloeden and Platen (1999) as well as by Rößler (2006). If the noise is additive, then it is possible to derive a Runge–Kutta counterpart of the method in Algorithm 5.8 which uses finite difference approximations instead of the closed-form derivatives (Kloeden et al., 1994).

Example 6.6 (Duffing van der Pol)**.** *Consider a simplified version of a Duffing van der Pol oscillator*

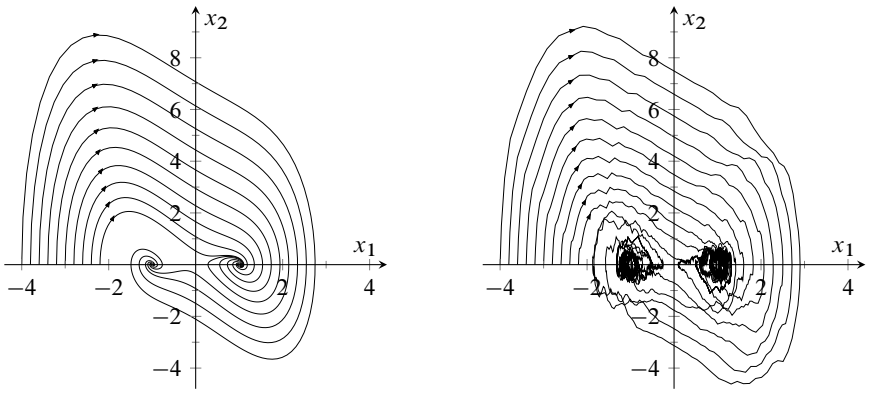
$$\ddot{x} + \dot{x} - (\alpha - x^2)\,x = x\,w(t), \quad \alpha \geq 0, \tag{6.36}$$

driven by multiplicative white noise $w(t)$ *with spectral density* q*. The corresponding two-dimensional,* $\mathbf{x}(t) = (x, \dot{x})$*, Itô stochastic differential equation is*

$$\begin{pmatrix} \mathrm{d}x_1 \\ \mathrm{d}x_2 \end{pmatrix} = \begin{pmatrix} x_2 \\ x_1\,(\alpha - x_1^2) - x_2 \end{pmatrix} \mathrm{d}t + \begin{pmatrix} 0 \\ x_1 \end{pmatrix} \mathrm{d}\beta, \tag{6.37}$$

where $\beta(t)$ *is a one-dimensional Brownian motion. The deterministic version (when* $q = 0$*), has the steady states* $\mathbf{x} = (0, 0)$ *and* $\mathbf{x} = (\pm\sqrt{\alpha}, 0)$*, the first of which is also a degenerate stationary state of the stochastic differential equation.*

Let $\alpha = 1$*. First we study the deterministic solution with no diffusion (*$q = 0$*). Figure 6.2 shows 10 trajectories, each with different initial values* $x_1(0)$*. We use*



(a) Deterministic **(b)** Stochastic

Figure 6.2: Trajectories simulated from the Duffing van der Pol oscillator model. The deterministic solutions ($q = 0$) converge to either of the two steady states. The realizations of the noise are identical for each of the stochastic ($q = 0.5^2$) trajectories.

a step size of $\Delta t = 2^{-5}$ *and a time-span of* $[0, 20]$. *We then replicate the result, but using the SDE model with* $q = 0.5^2$ *and identical realizations of noise in each trajectory. We use the strong order 1.0 method in Algorithm 6.3 for simulating the trajectories. Figure 6.3 shows the evolution of the trajectories.*

6.3 Weak stochastic Runge–Kutta methods

In the previous chapter we saw that it is possible to form weak approximations to SDEs, where the interest is not in the solution trajectories, but the distribution of them. It is often computationally convenient to replace the weak Itô–Taylor approximations by Runge–Kutta style approximations which avoid the use of derivatives of the drift and diffusion coefficients.

As an example of such a weak scheme, we consider the following scalar weak order 2.0 Runge–Kutta scheme for time-invariant SDEs due to Platen (see Kloeden and Platen, 1999), where the iteration takes the form:

$$\begin{aligned}\hat{x}(t_{k+1}) = \hat{x}(t_k) &+ \frac{1}{2}\big[f(\hat{x}(t_k)) + f(\tilde{x})\big]\Delta t \\ &+ \frac{1}{4}\big[L(\tilde{x}^+) + L(\tilde{x}^-) + 2\,L(\hat{x}_k)\big]\Delta\hat{\beta}_k \\ &+ \frac{1}{4\sqrt{\Delta t}}\big[L(\tilde{x}^+) + L(\tilde{x}^-)\big]\big[(\Delta\hat{\beta}_k)^2 + \Delta t\big] \qquad (6.38)\end{aligned}$$

with supporting values $\tilde{x} = \hat{x}_k + f(\hat{x}_k)\,\Delta t + L(\hat{x}_k)\,\Delta\hat{\beta}_k$ and $\tilde{x}^{\pm} = \hat{x}_k + f(\hat{x}_k)\,\Delta t \pm L(\hat{x}_k)\,\sqrt{\Delta t}$.

Rößler (2009) considers a general class of multi-dimensional weak order 2.0 stochastic Runge–Kutta schemes. The general multi-dimensional schemes by Rößler are as follows:

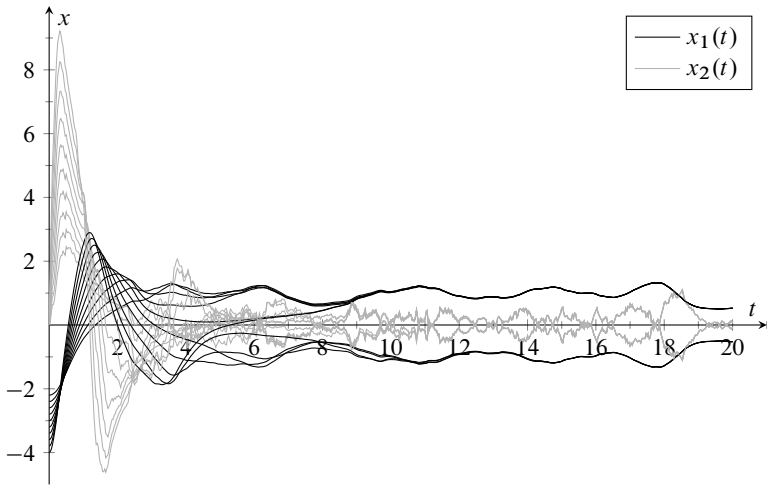


Figure 6.3: Evolution of the trajectories in the stochastic Duffing van der Pol oscillator model in Figure 6.2.

Algorithm 6.4 (A class of stochastic Runge–Kutta methods of weak order 2.0)**.** *Start from* $\hat{\mathbf{x}}(t_0) = \mathbf{x}(t_0)$ *and divide the integration interval* $[t_0, t]$ *into* n *steps* $t_0 < t_1 < t_2 < \ldots < t_n = t$ *such that* $\Delta t = t_{k+1} - t_k$*. The integration method is characterized by the following* extended Butcher tableau*:*

$$\begin{array}{c|c|cc} \mathbf{c}^{(0)} & \mathbf{A}^{(0)} & \mathbf{B}^{(0)} & \\ \hline \mathbf{c}^{(1)} & \mathbf{A}^{(1)} & \mathbf{B}^{(1)} & \\ \hline \mathbf{c}^{(2)} & \mathbf{A}^{(2)} & \mathbf{B}^{(2)} & \\ \hline & \boldsymbol{\alpha}^{\mathsf{T}} & \left[\boldsymbol{\gamma}^{(1)}\right]^{\mathsf{T}} & \left[\boldsymbol{\gamma}^{(2)}\right]^{\mathsf{T}} \\ \cline{2-4} & & \left[\boldsymbol{\gamma}^{(3)}\right]^{\mathsf{T}} & \left[\boldsymbol{\gamma}^{(4)}\right]^{\mathsf{T}} \end{array} \tag{6.39}$$

On each step k *approximate the solution by the following:*

$$\begin{aligned} \hat{\mathbf{x}}(t_{k+1}) = \hat{\mathbf{x}}(t_k) &+ \sum_{i=1}^{s} \alpha_i \, \mathbf{f}\left(\tilde{\mathbf{x}}_i^{(0)}, t_k + c_i^{(0)} \Delta t\right) \Delta t \\ &+ \sum_{i=1}^{s} \sum_{n=1}^{m} \gamma_i^{(1)} \, \mathbf{L}^n\left(\tilde{\mathbf{x}}_i^{(n)}, t_k + c_i^{(1)} \Delta t\right) \Delta \hat{\beta}_k^{(n)} \\ &+ \sum_{i=1}^{s} \sum_{n=1}^{m} \gamma_i^{(2)} \, \mathbf{L}^n\left(\tilde{\mathbf{x}}_i^{(n)}, t_k + c_i^{(1)} \Delta t\right) \frac{\Delta \hat{\beta}_k^{(n,n)}}{\sqrt{\Delta t}} \\ &+ \sum_{i=1}^{s} \sum_{n=1}^{m} \gamma_i^{(3)} \, \mathbf{L}^n\left(\bar{\mathbf{x}}_i^{(n)}, t_k + c_i^{(2)} \Delta t\right) \Delta \hat{\beta}_k^{(n)} \end{aligned}$$

$$+ \sum_{i=1}^{s} \sum_{n=1}^{m} \gamma_i^{(4)} \, \mathbf{L}^n(\bar{\mathbf{x}}_i^{(n)}, t_k + c_i^{(2)} \Delta t) \, \sqrt{\Delta t}, \tag{6.40}$$

with supporting values

$$\begin{aligned}
\tilde{\mathbf{x}}_i^{(0)} &= \hat{\mathbf{x}}(t_k) + \sum_{j=1}^{s} A_{i,j}^{(0)} \, \mathbf{f}(\tilde{\mathbf{x}}_j^{(0)}, t_k + c_j^{(0)} \Delta t) \, \Delta t \\
&\quad + \sum_{j=1}^{s} \sum_{l=1}^{m} B_{i,j}^{(0)} \, \mathbf{L}^l(\tilde{\mathbf{x}}_j^{(l)}, t_k + c_j^{(1)} \Delta t) \, \Delta \hat{\beta}_k^{(l)}, \qquad (6.41) \\
\tilde{\mathbf{x}}_i^{(n)} &= \hat{\mathbf{x}}(t_k) + \sum_{j=1}^{s} A_{i,j}^{(1)} \, \mathbf{f}(\tilde{\mathbf{x}}_j^{(0)}, t_k + c_j^{(0)} \Delta t) \, \Delta t \\
&\quad + \sum_{j=1}^{s} \sum_{l=1}^{m} B_{i,j}^{(1)} \, \mathbf{L}^l(\tilde{\mathbf{x}}_j^{(l)}, t_k + c_j^{(1)} \Delta t) \, \Delta \hat{\beta}_k^{(l,n)}, \qquad (6.42) \\
\bar{\mathbf{x}}_i^{(n)} &= \hat{\mathbf{x}}(t_k) + \sum_{j=1}^{s} A_{i,j}^{(2)} \, \mathbf{f}(\tilde{\mathbf{x}}_j^{(0)}, t_k + c_j^{(0)} \Delta t) \, \Delta t \\
&\quad + \sum_{j=1}^{s} \sum_{\substack{l=1 \\ l \neq n}}^{m} B_{i,j}^{(2)} \, \mathbf{L}^l(\tilde{\mathbf{x}}_j^{(l)}, t_k + c_j^{(1)} \Delta t) \, \frac{\Delta \hat{\beta}_k^{(l,n)}}{\sqrt{\Delta t}}, \qquad (6.43)
\end{aligned}$$

for $i = 1, 2, \ldots, s$ *and* $n = 1, 2, \ldots, m$.

The increments in the above algorithm are given by the double Itô integrals (exactly as in the case of the strong stochastic Runge–Kutta schemes), but in the weak schemes we can use the following approximations (see, *e.g.*, Kloeden and Platen, 1999; Rößler, 2009):

$$\Delta \hat{\beta}_k^{(i,j)} = \begin{cases} \frac{1}{2}\big(\Delta \hat{\beta}_k^{(i)} \, \Delta \hat{\beta}_k^{(j)} - \sqrt{\Delta t} \, \hat{\xi}_k^{(i)}\big), & \text{if } i < j, \\ \frac{1}{2}\big(\Delta \hat{\beta}_k^{(i)} \, \Delta \hat{\beta}_k^{(j)} + \sqrt{\Delta t} \, \hat{\xi}_k^{(j)}\big), & \text{if } i > j, \\ \frac{1}{2}\big([\Delta \hat{\beta}_k^{(i)}]^2 - \Delta t\big), & \text{if } i = j, \end{cases} \tag{6.44}$$

for $i, j = 1, 2, \ldots, m$. Here only $2m-1$ independent random variables are needed, and we do not anymore run into problems with the cross-term integrals as we did in the strong stochastic Runge–Kutta schemes. For example, we can choose $\Delta \hat{\beta}_k^{(i)}$ such that they are independent three-point distributed random variables

$$\mathrm{P}\big(\Delta \hat{\beta}_k^{(i)} = \pm \sqrt{3 \, \Delta t}\big) = \frac{1}{6} \quad \text{and} \quad \mathrm{P}\big(\Delta \hat{\beta}_k^{(i)} = 0\big) = \frac{2}{3}, \tag{6.45}$$

and the supporting variables $\hat{\zeta}_k^{(i)}$ such that they are independent two-point distributed random variables

$$\mathrm{P}\big(\hat{\zeta}_k^{(i)} = \pm\sqrt{\Delta t}\big) = \frac{1}{2}. \tag{6.46}$$

Rößler (2009) proposes, for example, the following multi-dimensional weak order 2.0 stochastic Runge–Kutta scheme, which only requires two evaluations of $\mathbf{f}(\cdot,\cdot)$ and only five evaluations of each $\mathbf{L}^i(\cdot,\cdot)$.

Algorithm 6.5 (Weak order 2.0 stochastic Runge–Kutta due to Rößler)**.** *Consider a stochastic Runge–Kutta method with the following extended Butcher tableau:*

$$\begin{array}{c|cc|cc|cc}
0 & & & & & & \\
1 & 1 & & 1 & & & \\
0 & 0 & 0 & 0 & 0 & & \\
\hline
0 & & & & & & \\
1 & 1 & & 1 & & & \\
1 & 1 & 0 & -1 & 0 & & \\
\hline
0 & & & & & & \\
1 & 1 & & 1 & & & \\
1 & 1 & 0 & -1 & 0 & & \\
\hline
 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \quad \frac{1}{2} \quad -\frac{1}{2} \\
\hline
 & & & -\frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \quad \frac{1}{2} \quad -\frac{1}{2}
\end{array} \tag{6.47}$$

which corresponds to the following iterative scheme

$$\begin{aligned}
\hat{\mathbf{x}}(t_{k+1}) = \hat{\mathbf{x}}(t_k) &+ \frac{\Delta t}{2}\Big[\mathbf{f}(\hat{\mathbf{x}}(t_k), t_k) + \mathbf{f}(\tilde{\mathbf{x}}_2^{(0)}, t_k + \Delta t)\Big] \\
&+ \sum_{n=1}^{m}\left[\frac{1}{2}\mathbf{L}^n(\hat{\mathbf{x}}(t_k), t_k) + \frac{1}{4}\mathbf{L}^n(\tilde{\mathbf{x}}_2^{(n)}, t_k + \Delta t) + \frac{1}{4}\mathbf{L}^n(\tilde{\mathbf{x}}_3^{(n)}, t_k + \Delta t)\right]\Delta\hat{\beta}_k^{(n)} \\
&+ \sum_{n=1}^{m}\left[\frac{1}{2}\mathbf{L}^n(\tilde{\mathbf{x}}_2^{(n)}, t_k + \Delta t) - \frac{1}{2}\mathbf{L}^n(\tilde{\mathbf{x}}_3^{(n)}, t_k + \Delta t)\right]\frac{\Delta\hat{\beta}_k^{(n,n)}}{\sqrt{\Delta t}} \\
&+ \sum_{n=1}^{m}\left[-\frac{1}{2}\mathbf{L}^n(\hat{\mathbf{x}}(t_k), t_k) + \frac{1}{4}\mathbf{L}^n(\bar{\mathbf{x}}_2^{(n)}, t_k + \Delta t) + \frac{1}{4}\mathbf{L}^n(\bar{\mathbf{x}}_3^{(n)}, t_k + \Delta t)\right]\Delta\hat{\beta}_k^{(n)} \\
&+ \sum_{n=1}^{m}\left[\frac{1}{2}\mathbf{L}^n(\bar{\mathbf{x}}_2^{(n)}, t_k + \Delta t) - \frac{1}{2}\mathbf{L}^n(\bar{\mathbf{x}}_3^{(n)}, t_k + \Delta t)\right]\sqrt{\Delta t}
\end{aligned} \tag{6.48}$$

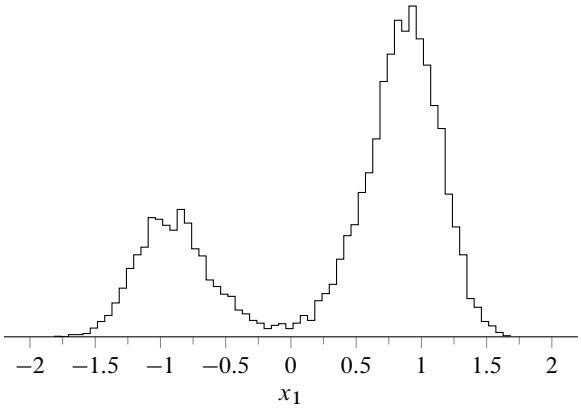


Figure 6.4: A histogram of the state of the Duffing van der Pol oscillator values $x_1(t)$ at $t = 20$ with 10,000 samples simulated by the weak order 2.0 stochastic Runge–Kutta scheme.

with supporting values (note that $\tilde{\mathbf{x}}_1^{(0)} = \tilde{\mathbf{x}}_1^{(n)} = \bar{\mathbf{x}}_1^{(n)} = \hat{\mathbf{x}}(t_k)$*)*

$$\tilde{\mathbf{x}}_2^{(0)} = \hat{\mathbf{x}}(t_k) + \mathbf{f}\left(\hat{\mathbf{x}}(t_k), t_k\right) \Delta t + \sum_{l=1}^{m} \mathbf{L}^l\left(\hat{\mathbf{x}}(t_k), t_k\right) \Delta\hat{\beta}_k^{(l)}, \tag{6.49}$$

$$\tilde{\mathbf{x}}_2^{(n)} = \hat{\mathbf{x}}(t_k) + \mathbf{f}\left(\hat{\mathbf{x}}(t_k), t_k\right) \Delta t + \sum_{l=1}^{m} \mathbf{L}^l\left(\hat{\mathbf{x}}(t_k), t_k\right) \Delta\hat{\beta}_k^{(l,n)}, \tag{6.50}$$

$$\tilde{\mathbf{x}}_3^{(n)} = \hat{\mathbf{x}}(t_k) + \mathbf{f}\left(\hat{\mathbf{x}}(t_k), t_k\right) \Delta t - \sum_{l=1}^{m} \mathbf{L}^l\left(\hat{\mathbf{x}}(t_k), t_k\right) \Delta\hat{\beta}_k^{(l,n)}, \tag{6.51}$$

$$\bar{\mathbf{x}}_2^{(n)} = \hat{\mathbf{x}}(t_k) + \mathbf{f}\left(\hat{\mathbf{x}}(t_k), t_k\right) \Delta t + \sum_{\substack{l=1 \\ l\neq n}}^{m} \mathbf{L}^l\left(\hat{\mathbf{x}}(t_k), t_k\right) \frac{\Delta\hat{\beta}_k^{(l,n)}}{\sqrt{\Delta t}}, \tag{6.52}$$

$$\bar{\mathbf{x}}_3^{(n)} = \hat{\mathbf{x}}(t_k) + \mathbf{f}\left(\hat{\mathbf{x}}(t_k), t_k\right) \Delta t - \sum_{\substack{l=1 \\ l\neq n}}^{m} \mathbf{L}^l\left(\hat{\mathbf{x}}(t_k), t_k\right) \frac{\Delta\hat{\beta}_k^{(l,n)}}{\sqrt{\Delta t}}. \tag{6.53}$$

Example 6.7 (Weak approximation of the Duffing van der Pol problem)**.** *In Example 6.6 we considered a van der Pol oscillator with two steady states for the zero-diffusion model. Now we are interested in characterizing the solution at* $t = 20$ *for the initial condition of* $\mathbf{x}(0) = (-3, 0)$*. We use the stochastic Runge–Kutta method in Algorithm 6.5 that is of weak order 2.0. We consider a time-span* $[0, 20]$ *and a discretization interval* $\Delta t = 2^{-4}$*. With a* Δt *this large, the Euler–Maruyama method does not provide plausible results. Figure 6.4 shows the histogram of the values* $x_1(20)$*.*

Chapter 7

Bayesian estimation of SDEs

7.1 Bayesian filtering in SDE models

Filtering theory (*e.g.*, Stratonovich, 1968; Jazwinski, 1970; Maybeck, 1979, 1982; Särkkä, 2006; Crisan and Rozovskii, 2011; Särkkä, 2013) is introduced in this material with the following problem. Assume that we have a pair of processes $(\mathbf{x}(t), \mathbf{y}(t))$ such that $\mathbf{y}(t)$ is observed and $\mathbf{x}(t)$ is hidden. Now the question is: Given that we have observed $\mathbf{y}(t)$, what can we say (in statistical sense) about the hidden process $\mathbf{x}(t)$? In particular, the main question in Bayesian sense is what is the conditional probability distribution of the hidden process $\mathbf{x}(t)$ given the observed process $\mathbf{y}(t)$.

Example 7.1 (Continuous-time car tracking model)**.** *Recall that in Example 2.5 we modeled the dynamics of a car via the white-noise-force Newton's law*

$$\frac{\mathrm{d}^2 x_1}{\mathrm{d}t^2} = w_1(t),$$
$$\frac{\mathrm{d}^2 x_2}{\mathrm{d}t^2} = w_2(t).$$

(see Figure 7.1a) which then resulted in a SDE model of the form (in white noise interpretation):

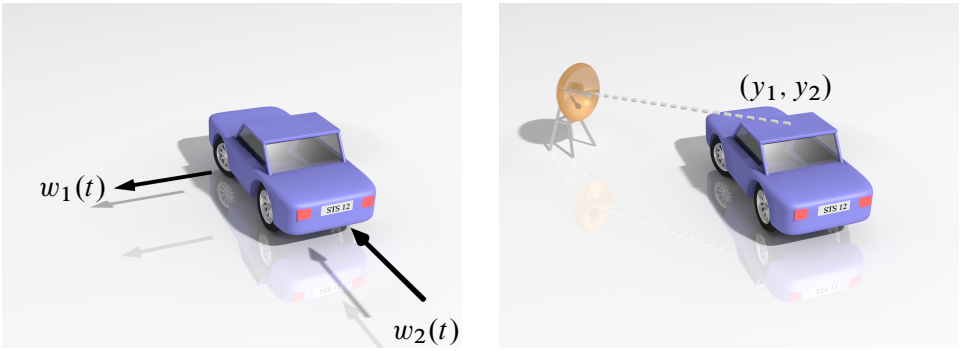
$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{F}\,\mathbf{x} + \mathbf{L}\,\mathbf{w}.$$

Let's now assume that we use a radar to obtain noisy measurements (y_1, y_2) of the car's position, which thus can be modeled as

$$y_1(t) = x_1(t) + \varepsilon_1(t),$$
$$y_2(t) = x_2(t) + \varepsilon_2(t),$$

where $\varepsilon_1(t)$ and $\varepsilon_2(t)$ are white noise processes (see Figure 7.1b). The measurement model can now be written as

$$\mathbf{y}(t) = \mathbf{H}\,\mathbf{x}(t) + \boldsymbol{\varepsilon}(t), \qquad \mathbf{H} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}. \tag{7.1}$$



(a) Dynamic model **(b)** Measurement model

Figure 7.1: Illustration of **(a)** a dynamic and **(b)** a measurement model of a car. In the dynamic model, the unknown forces $w_1(t)$ and $w_2(t)$ are modeled as white noise processes. The measurements (y_1, y_2) are modeled as noise corrupted observations of the car's position.

If we now interpret both the dynamic model and measurement model as proper SDEs, the resulting model can be written as

$$\begin{aligned} \mathrm{d}\mathbf{x} &= \mathbf{F}\,\mathbf{x}\,\mathrm{d}t + \mathbf{L}\,\mathrm{d}\boldsymbol{\beta}, \\ \mathrm{d}\mathbf{z} &= \mathbf{H}\,\mathbf{x}\,\mathrm{d}t + \mathrm{d}\boldsymbol{\eta}, \end{aligned} \tag{7.2}$$

where formally $\mathbf{y} = \mathrm{d}\mathbf{z}/\mathrm{d}t$ *and* $\boldsymbol{\varepsilon} = \mathrm{d}\boldsymbol{\eta}/\mathrm{d}t$.

The model in Equation (7.2) is a canonical example of a continuous-time filtering model. The corresponding *Bayesian filtering problem* on this model is to determine the conditional distribution of the "state" $\mathbf{x}(t)$ given the history of measurements $\{\mathbf{z}(\tau) \mid 0 \leq \tau \leq t\}$—or equivalently given $\{\mathbf{y}(\tau) \mid 0 \leq \tau \leq t\}$. In the car example above this corresponds to determining the conditional distribution car position and velocity given the history of observations obtained so far. Given the conditional distribution, we can, for example, compute the conditional mean of the state $\mathbf{x}(t)$, which is also its minimum mean squared estimate, as well as its variance, which measures the accuracy of the estimate.

The above problem is a continuous-time filtering problem, because both the dynamics and measurements are modeled as continuous-time processes. However, from an applications point of view we often obtain measurements not as a continuous function, but at certain sampling times $t_1, t_2, \ldots, t_k$. In that case it is more convenient to formulate the problem as a continuous/discrete-time filtering problem, where the dynamics are continuous and the measurement process is in discrete time. This kind of continuous/discrete-time filtering problem are closely related to discrete-time filtering problems where both the dynamics are measurements are modeled in discrete time. The classical reference of continuous and continuous/discrete-time problems is the book of Jazwinski (1970) whereas a more

modern treatment of the subject can be found in Särkkä (2006). A quite comprehensive introduction to discrete-time filtering can be found in Särkkä (2013).

In the following sections we start by introducing the continuous-time filtering problem and then proceed to the discretely measured filtering problems.

7.2 Kushner–Stratonovich and Zakai equations, and Kalman–Bucy filtering

In mathematical terms a continuous-time filtering model can be written as

$$\begin{aligned} \mathrm{d}\mathbf{x}(t) &= \mathbf{f}(\mathbf{x}(t), t)\,\mathrm{d}t + \mathbf{L}(\mathbf{x}, t)\,\mathrm{d}\boldsymbol{\beta}(t), \\ \mathrm{d}\mathbf{z}(t) &= \mathbf{h}(\mathbf{x}(t), t)\,\mathrm{d}t + \mathrm{d}\boldsymbol{\eta}(t), \end{aligned} \tag{7.3}$$

where the first equation is the dynamic model and second the measurement model. In the equation we have the following:

- $\mathbf{x}(t) \in \mathbb{R}^n$ is the state process,
- $\mathbf{z}(t) \in \mathbb{R}^m$ is the (integrated) measurement process,
- $\mathbf{f}$ is the drift function,
- $\mathbf{h}$ is the measurement model function,
- $\mathbf{L}(\mathbf{x}, t)$ is the dispersion matrix,
- $\boldsymbol{\beta}(t)$ and $\boldsymbol{\eta}(t)$ are independent Brownian motions with diffusion matrices $\mathbf{Q}$ and $\mathbf{R}$, respectively.

In physical sense the measurement model is easier to understand by writing it in white noise form:

$$\mathbf{y}(t) = \mathbf{h}(\mathbf{x}(t), t) + \boldsymbol{\varepsilon}(t), \tag{7.4}$$

where we have defined the physical measurement as $\mathbf{y}(t) = \mathrm{d}\mathbf{z}(t)/\mathrm{d}t$, and $\boldsymbol{\varepsilon}(t) = \mathrm{d}\boldsymbol{\eta}(t)/\mathrm{d}t$ is the formal derivative of $\boldsymbol{\eta}(t)$. That is, we measure the state through the non-linear measurement model $h(\bullet)$, and the measurement is corrupted with continuous-time white noise $\boldsymbol{\varepsilon}(t)$. Typical applications of this kind of models are, for example, tracking and navigation problems, where $\mathbf{x}(t)$ is the dynamic state of the target—say, the position and velocity of a car. The measurement can be, for example, radar readings which contain some noise $\boldsymbol{\varepsilon}(t)$.

The most natural framework to formulate the problem of estimation of the state from the measurement is in terms of Bayesian inference. This is indeed the classical formulation of non-linear filtering theory and was used already in the books of Stratonovich (1968) and Jazwinski (1970). The purpose of the continuous-time optimal (Bayesian) filter is to compute the posterior distribution (or the filtering

distribution) of the process $\mathbf{x}(t)$ given the observed process (more precisely, the *sigma algebra* generated by the observed process)

$$\mathcal{Y}_t = \{\mathbf{y}(\tau) \,:\, 0 \leq \tau \leq t\} = \{\mathbf{z}(\tau) \,:\, 0 \leq \tau \leq t\}, \tag{7.5}$$

that is, we wish to compute

$$p(\mathbf{x}(t) \mid \mathcal{Y}_t). \tag{7.6}$$

In the following we present the general equations for computing these distributions. To ease the notation let us recall the following definition of the operator occurring in the Fokker–Planck–Kolmogorov equation, which was introduced in Equation (4.14):

$$\begin{aligned} \mathcal{A}^*(\bullet) = &-\sum_i \frac{\partial}{\partial x_i}[f_i(x,t)\,(\bullet)] \\ &+ \frac{1}{2}\sum_{ij} \frac{\partial^2}{\partial x_i\,\partial x_j}\{[\mathbf{L}(\mathbf{x},t)\,\mathbf{Q}\,\mathbf{L}^\mathsf{T}(\mathbf{x},t)]_{ij}\,(\bullet)\}, \end{aligned} \tag{7.7}$$

which thus allows us to write the Fokker–Planck–Kolmogorov equation (4.2) compactly as

$$\frac{\partial p}{\partial t} = \mathcal{A}^* p. \tag{7.8}$$

The continuous-time optimal filtering equation, which computes $p(\mathbf{x}(t) \mid \mathcal{Y}_t)$ is called the Kushner–Stratonovich (KS) equation (Kushner, 1964; Bucy, 1965) and can be derived as the continuous-time limits of the so called Bayesian filtering equations (see, *e.g.*, Särkkä, 2013). A Stratonovich calculus version of the equation was studied by Stratonovich already in late 1950's (*cf.* Stratonovich, 1968).

Algorithm 7.1 (Kushner–Stratonovich equation)**.** *The stochastic partial differential equation for the filtering density* $p(\mathbf{x}, t \mid \mathcal{Y}_t) \triangleq p(\mathbf{x}(t) \mid \mathcal{Y}_t)$ *is*

$$\begin{aligned} \mathrm{d}p(\mathbf{x}, t \mid \mathcal{Y}_t) = &\mathcal{A}^*\, p(\mathbf{x}, t \mid \mathcal{Y}_t)\,\mathrm{d}t \\ &+ (\mathbf{h}(\mathbf{x},t) - \mathrm{E}[\mathbf{h}(\mathbf{x},t) \mid \mathcal{Y}_t])^\mathsf{T}\,\mathbf{R}^{-1}(\mathrm{d}\mathbf{z} - \mathrm{E}[\mathbf{h}(\mathbf{x},t) \mid \mathcal{Y}_t]\,\mathrm{d}t)\; p(\mathbf{x}, t \mid \mathcal{Y}_t), \end{aligned} \tag{7.9}$$

where $\mathrm{d}p(\mathbf{x}, t \mid \mathcal{Y}_t) = p(\mathbf{x}, t + \mathrm{d}t \mid \mathcal{Y}_{t+\mathrm{d}t}) - p(\mathbf{x}, t \mid \mathcal{Y}_t)$ *and*

$$\mathrm{E}[\mathbf{h}(\mathbf{x},t) \mid \mathcal{Y}_t] = \int \mathbf{h}(\mathbf{x},t)\; p(\mathbf{x}, t \mid \mathcal{Y}_t)\,\mathrm{d}\mathbf{x}. \tag{7.10}$$

This equation is only formal in the sense that as such it is quite much impossible to work with. However, it is possible derive all kinds of moment equations from it, as well as form approximations to the solutions. What makes the equation difficult is that it is a non-linear stochastic partial differential equation—recall that the operator $\mathcal{A}^*$ contains partial derivatives. Furthermore the equation is non-linear, as could be seen by expanding the expectation integrals in the equation (recall that

they are integrals over $p(\mathbf{x}, t \mid \mathcal{Y}_t)$). The stochasticity is generated by the observation process $\mathbf{z}(t)$.

The nonlinearity in the KS equation can be eliminated by deriving an equation for an unnormalized filtering distribution instead of the normalized one. This leads to so called Zakai equation (Zakai, 1969).

Algorithm 7.2 (Zakai equation)**.** *Let* $q(\mathbf{x}, t \mid \mathcal{Y}_t) \triangleq q(\mathbf{x}(t) \mid \mathcal{Y}_t)$ *be the solution to Zakai's stochastic partial differential equation*

$$\mathrm{d}q(\mathbf{x}, t \mid \mathcal{Y}_t) = \mathcal{A}^* \, q(\mathbf{x}, t \mid \mathcal{Y}_t) \, \mathrm{d}t + \mathbf{h}^\mathsf{T}(\mathbf{x}, t) \, \mathbf{R}^{-1} \, \mathrm{d}\mathbf{z} \, q(\mathbf{x}, t \mid \mathcal{Y}_t), \tag{7.11}$$

where $\mathrm{d}q(\mathbf{x}, t \mid \mathcal{Y}_t) = q(\mathbf{x}, t + \mathrm{d}t \mid \mathcal{Y}_{t+\mathrm{d}t}) - q(\mathbf{x}, t \mid \mathcal{Y}_t)$ *and* $\mathcal{A}^*$ *is the Fokker–Planck–Kolmogorov operator defined in Equation* (4.14)*. Then we have*

$$p(\mathbf{x}(t) \mid \mathcal{Y}_t) = \frac{q(\mathbf{x}(t) \mid \mathcal{Y}_t)}{\int q(\mathbf{x}(t) \mid \mathcal{Y}_t) \, \mathrm{d}\mathbf{x}(t)}. \tag{7.12}$$

The car model in Example 7.1 was actually a linear Gaussian filtering problem, which refer to a problem where the functions $\mathbf{f}$ and $\mathbf{h}$ are linear in $\mathbf{x}$. In that case the filtering solution is Gaussian and we can solve the filtering equations in closed form. The *Kalman–Bucy filter* (Kalman and Bucy, 1961) is the exact solution to the linear Gaussian filtering problem

$$\begin{aligned} \mathrm{d}\mathbf{x} &= \mathbf{F}(t) \, \mathbf{x} \, \mathrm{d}t + \mathbf{L}(t) \, \mathrm{d}\boldsymbol{\beta}, \\ \mathrm{d}\mathbf{z} &= \mathbf{H}(t) \, \mathbf{x} \, \mathrm{d}t + \mathrm{d}\boldsymbol{\eta}, \end{aligned} \tag{7.13}$$

where

- $\mathbf{x}(t) \in \mathbb{R}^n$ is the state process,
- $\mathbf{z}(t) \in \mathbb{R}^m$ is the (integrated) measurement process,
- $\mathbf{F}(t)$ is the dynamic model matrix,
- $\mathbf{H}(t)$ is the measurement model matrix,
- $\mathbf{L}(t)$ is an arbitrary time varying matrix, independent of $\mathbf{x}(t)$ and $\mathbf{y}(t)$,
- $\boldsymbol{\beta}(t)$ and $\boldsymbol{\eta}(t)$ are independent Brownian motions with diffusion matrices $\mathbf{Q}$ and $\mathbf{R}$, respectively.

The solution is given as follows:

Algorithm 7.3 (Kalman–Bucy filter)**.** *The Bayesian filter, which computes the posterior distribution* $p(\mathbf{x}(t) \mid \mathcal{Y}_t) = \mathrm{N}(\mathbf{x}(t) \mid \mathbf{m}(t), \mathbf{P}(t))$ *for the system* (7.13) *is*

$$\begin{aligned} \mathbf{K}(t) &= \mathbf{P}(t) \, \mathbf{H}^\mathsf{T}(t) \, \mathbf{R}^{-1}, \\ \mathrm{d}\mathbf{m}(t) &= \mathbf{F}(t) \, \mathbf{m}(t) \, \mathrm{d}t + \mathbf{K}(t) \, [\mathrm{d}\mathbf{z}(t) - \mathbf{H}(t) \, \mathbf{m}(t) \, \mathrm{d}t], \\ \frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t} &= \mathbf{F}(t) \, \mathbf{P}(t) + \mathbf{P}(t) \, \mathbf{F}^\mathsf{T}(t) + \mathbf{L}(t) \, \mathbf{Q} \, \mathbf{L}^\mathsf{T}(t) - \mathbf{K}(t) \, \mathbf{R} \, \mathbf{K}^\mathsf{T}(t). \end{aligned} \tag{7.14}$$

7.3 Continuous-time approximate non-linear filtering

There exists various approximation methods so cope with non-linear models, for example, based on Monte Carlo approximations, series expansions of processes and densities, Gaussian (process) approximations and many others (see, *e.g.*, Crisan and Rozovskii, 2011). One way which is utilized in many practical applications is to use Gaussian approximations outlined in the beginning of Chapter 5. The classical filtering theory is very much based on this idea and a typical approach is to use Taylor series expansions of the drift function (Jazwinski, 1970). The use of Gaussian sigma-point type of approximations in this context has been recently studied in Särkkä (2007) and Särkkä and Sarmavuori (2013). In this section we only outline Gaussian approximation based approximate filtering and for other methods reader is referred to Crisan and Rozovskii (2011).

The extended Kalman–Bucy filter (see, *e.g.*, Gelb, 1974) is perhaps the most common and the simplest possible extension of the Kalman–Bucy filter to non-linear models of the form (7.3). It can be derived by using a first order Taylor series expansions on the functions $\mathbf{f}$ and $\mathbf{h}$ around the current mean estimate.

Algorithm 7.4 (Extended Kalman–Bucy filter)**.** *The equations of the extended Kalman–Bucy filter (EKBF) are:*

$$\begin{aligned}
\mathbf{K}(t) &= \mathbf{P}(t)\,\mathbf{H}^\mathsf{T}(\mathbf{m}(t),t)\,\mathbf{R}^{-1} \\
\mathrm{d}\mathbf{m}(t) &= \mathbf{f}(\mathbf{m}(t),t)\,\mathrm{d}t + \mathbf{K}(t)\,[\mathrm{d}\mathbf{z}(t) - \mathbf{h}(\mathbf{m}(t),t)\,\mathrm{d}t] \\
\frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t} &= \mathbf{F}(\mathbf{m}(t),t)\,\mathbf{P}(t) + \mathbf{P}(t)\,\mathbf{F}^\mathsf{T}(\mathbf{m}(t),t) \\
&\quad + \mathbf{L}(\mathbf{m}(t),t)\,\mathbf{Q}\,\mathbf{L}^\mathsf{T}(\mathbf{m}(t),t) - \mathbf{K}(t)\,\mathbf{R}\,\mathbf{K}^\mathsf{T}(t),
\end{aligned} \tag{7.15}$$

where $\mathbf{F}$ *is the Jacobian matrix of* $\mathbf{f}$ *with elements* $F_{ij} = \partial f_i/\partial x_j$*, and* $\mathbf{H}$ *is the Jacobian matrix of* $\mathbf{h}$ *with elements* $H_{ij} = \partial h_i/\partial x_j$*.*

It is now easy to see that we have actually employed the linearization approximation from Algorithm 5.3 here. Taking a step backwards lets us now use Algorithm 5.1 to formulate the following general Gaussian approximation to the non-linear filtering problem (see, *e.g.*, Särkkä and Sarmavuori, 2013).

Algorithm 7.5 (Continuous-time Gaussian filter)**.** *The equations of the continuous-time Gaussian filter are:*

$$\begin{aligned}
\mathbf{K}(t) &= \mathrm{E}_\mathrm{N}[(\mathbf{x} - \mathbf{m}(t))\,\mathbf{h}^\mathsf{T}(\mathbf{x}(t),t)]\,\mathbf{R}^{-1}, \\
\mathrm{d}\mathbf{m} &= \mathrm{E}_\mathrm{N}[\mathbf{f}(\mathbf{x},t)]\,\mathrm{d}t + \mathbf{K}(t)\,(\mathrm{d}\mathbf{z} - \mathrm{E}_\mathrm{N}[\mathbf{h}(\mathbf{x},t)]\,\mathrm{d}t), \\
\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}t} &= \mathrm{E}_\mathrm{N}[(\mathbf{x} - \mathbf{m}(t))\,\mathbf{f}^\mathsf{T}(\mathbf{x},t)] + \mathrm{E}_\mathrm{N}[\mathbf{f}(\mathbf{x},t)\,(\mathbf{x} - \mathbf{m}(t))^\mathsf{T}] \\
&\quad + \mathrm{E}_\mathrm{N}[\mathbf{L}(\mathbf{x},t)\,\mathbf{Q}\,\mathbf{L}^\mathsf{T}(\mathbf{x},t)] - \mathbf{K}(t)\,\mathbf{R}\,\mathbf{K}^\mathsf{T}(t),
\end{aligned} \tag{7.16}$$

where the expectations are taken with respect to $\mathbf{x} \sim \mathrm{N}(\mathbf{m}(t), \mathbf{P}(t))$*. Using the integration by parts, we can write the above equations in an analogous form to*

Algorithm 5.2 as follows:

$$\begin{aligned}
\mathbf{K}(t) &= \mathbf{P}(t)\ \mathrm{E_N}[\mathbf{H}^\mathsf{T}(\mathbf{x},t)]\,\mathbf{R}^{-1},\\
\mathrm{d}\mathbf{m} &= \mathrm{E_N}[\mathbf{f}(\mathbf{x},t)]\,\mathrm{d}t + \mathbf{K}(t)\,(\mathrm{d}\mathbf{z} - \mathrm{E_N}[\mathbf{h}(\mathbf{x},t)]\,\mathrm{d}t),\\
\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}t} &= \mathrm{E_N}[\mathbf{F}(\mathbf{x},t)]\,\mathbf{P}(t) + \mathbf{P}(t)\ \mathrm{E_N}[\mathbf{F}^\mathsf{T}(\mathbf{x},t)]\\
&\quad + \mathrm{E_N}[\mathbf{L}(\mathbf{x},t)\,\mathbf{Q}\,\mathbf{L}^\mathsf{T}(\mathbf{x},t)] - \mathbf{K}(t)\,\mathbf{R}\,\mathbf{K}^\mathsf{T}(t).
\end{aligned} \tag{7.17}$$

Various sigma-point approximations to the continuous-time filtering problem can now be generated by replacing the Gaussian expectations above with sigma-point approximations analogously to Algorithm 5.4. The generic form resulting from approximating Equations (7.16) is the following.

Algorithm 7.6 (Continuous-time sigma-point filter)**.** *The equations of a generic continuous-time sigma-point filter are:*

$$\begin{aligned}
\mathbf{K}(t) &= \sum_i W^{(i)}\,\sqrt{\mathbf{P}}\,\boldsymbol{\xi}_i\,\mathbf{h}^\mathsf{T}(\mathbf{m} + \sqrt{\mathbf{P}}\,\boldsymbol{\xi}_i, t)\,\mathbf{R}^{-1},\\
\mathrm{d}\mathbf{m} &= \sum_i W^{(i)}\,\mathbf{f}(\mathbf{m} + \sqrt{\mathbf{P}}\,\boldsymbol{\xi}_i, t)\,\mathrm{d}t\\
&\quad + \mathbf{K}(t)\left(\mathrm{d}\mathbf{z} - \sum_i W^{(i)}\,\mathbf{h}(\mathbf{m} + \sqrt{\mathbf{P}}\,\boldsymbol{\xi}_i, t)\,\mathrm{d}t\right),\\
\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}t} &= \sum_i W^{(i)}\,\mathbf{f}(\mathbf{m} + \sqrt{\mathbf{P}}\,\boldsymbol{\xi}_i, t)\,\boldsymbol{\xi}_i^\mathsf{T}\,\sqrt{\mathbf{P}}^\mathsf{T}\\
&\quad + \sum_i W^{(i)}\,\sqrt{\mathbf{P}}\,\boldsymbol{\xi}_i\,\mathbf{f}^\mathsf{T}(\mathbf{m} + \sqrt{\mathbf{P}}\,\boldsymbol{\xi}_i, t)\\
&\quad + \sum_i W^{(i)}\,\mathbf{L}(\mathbf{m} + \sqrt{\mathbf{P}}\,\boldsymbol{\xi}_i, t)\,\mathbf{Q}\,\mathbf{L}^\mathsf{T}(\mathbf{m} + \sqrt{\mathbf{P}}\,\boldsymbol{\xi}_i, t)\\
&\quad - \mathbf{K}(t)\,\mathbf{R}\,\mathbf{K}^\mathsf{T}(t).
\end{aligned} \tag{7.18}$$

For details on selection of sigma-points $\boldsymbol{\xi}_i$ and weights $W^{(i)}$, see Section 5.2. For example, by selecting the unscented transform sigma-points and weights we get the unscented Kalman–Bucy filter (Särkkä, 2007).

7.4 Continuous/discrete-time Bayesian and Kalman filtering

In applications involving digital computers and computer controlled sensors we do not usually obtain measurements in continuous-time, but we are only able to get samples from the underlying process at discrete instants of time. For this kind of models continuous/discrete-time formulation of the problem is more appropriate.

A general continuous/discrete-time filtering problem can be formulated as

$$\begin{aligned} \mathrm{d}\mathbf{x} &= \mathbf{f}(\mathbf{x}, t)\,\mathrm{d}t + \mathbf{L}(\mathbf{x}, t)\,\mathrm{d}\boldsymbol{\beta}(t), \\ \mathbf{y}_k &\sim p(\mathbf{y}_k \mid \mathbf{x}(t_k)), \end{aligned} \tag{7.19}$$

where

- $\mathbf{x}(t) \in \mathbb{R}^n$ is the state,
- $\mathbf{y}_k \in \mathbb{R}^m$ is the measurement obtained at time instance t_k.
- $\mathbf{f} : \mathbb{R}^n \times \mathbb{R}_+ \to \mathbb{R}^n$ is the drift function,
- $\mathbf{L}(\mathbf{x}, t) \in \mathbb{R}^{n \times s}$ is the dispersion matrix,
- $\boldsymbol{\beta}(t) \in \mathbb{R}^s$ is Brownian motion with diffusion matrix, $\mathbf{Q} \in \mathbb{R}^{s \times s}$,
- $p(\mathbf{y}_k \mid \mathbf{x}(t_k))$ is the measurement model, which defines the distribution (or likelihood) of the measurement $\mathbf{y}_k$ given the state $\mathbf{x}(t_k)$.

In practice, we often construct the measurement model $p(\mathbf{y}_k \mid \mathbf{x}(t_k))$ as a noise corrupted measurement of the form

$$\mathbf{y}_k = \mathbf{h}_k(\mathbf{x}(t_k)) + \mathbf{r}_k, \tag{7.20}$$

where $\mathbf{r}_k \sim \mathrm{N}(\mathbf{0}, \mathbf{R}_k)$ is a Gaussian measurement noise. This thus implies that our measurement model is given as

$$p(\mathbf{y}_k \mid \mathbf{x}(t_k)) = \mathrm{N}(\mathbf{y}_k \mid \mathbf{h}_k(\mathbf{x}(t_k)), \mathbf{R}_k). \tag{7.21}$$

The model function $\mathbf{h}_k$ can, for example, map the state to a position measurement or into distance and direction measurements, which is more typically the case in radar applications.

The corresponding Bayesian filtering problem is now to determine the distributions

$$p(\mathbf{x}(t_k) \mid \mathbf{y}_1, \ldots, \mathbf{y}_k), \tag{7.22}$$

which are thus the posterior distributions of the states at the measurements times t_k given the measurements obtained so far. A bit more generally, we might be interested in determining the distributions

$$p(\mathbf{x}(t) \mid \mathbf{y}_1, \ldots, \mathbf{y}_k), \qquad t \in [t_k, t_{k+1}), \tag{7.23}$$

which also give the distributions of the state between the last and the next measurement. However, here we will only consider the former distributions although the latter can be easily obtained from continuous/discrete filters as well (Jazwinski, 1970; Särkkä, 2006).

Example 7.2 (Continuous/discrete-time car tracking model)**.** *Assume that we are tracking a car as in Example 7.1 except that we obtain measurements at discrete time instants* $t_1, t_2, \ldots$ *The measurement model can now be written as*

$$\mathbf{y}_k = \mathbf{H}\,\mathbf{x}_k + \mathbf{r}_k, \qquad \mathbf{H} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}. \tag{7.24}$$

where $\mathbf{r}_k \sim \mathrm{N}(\mathbf{0}, \mathbf{R}_k)$*. The dynamic and measurement models now form a continuous/discrete-time model*

$$\begin{aligned} \mathrm{d}\mathbf{x} &= \mathbf{F}\,\mathbf{x}\,\mathrm{d}t + \mathbf{L}\,\mathrm{d}\boldsymbol{\beta}, \\ \mathbf{y}_k &= \mathbf{H}\,\mathbf{x}(t_k) + \mathbf{r}_k, \end{aligned}$$

and the car tracking problem reduces to computing the conditional distribution of the state $\mathbf{x}(t)$ *given the measurements* $\mathbf{y}_1, \mathbf{y}_2, \ldots$ *obtained before the time* t*. These distributions are exactly the distributions* (7.23) *above.*

A conceptually simple way of dealing with the continuous/discrete-time filtering problem is to solve the transition densities $p(\mathbf{x}(t_{k+1}) \mid \mathbf{x}(t_k))$ from the Fokker–Planck–Kolmogorov forward partial differential equation (see Section 4.3, Theorem 4.2). The filtering problem takes the form

$$\begin{aligned} \mathbf{x}(t_{k+1}) &\sim p(\mathbf{x}(t_{k+1}) \mid \mathbf{x}(t_k)), \\ \mathbf{y}_k &\sim p(\mathbf{y}_k \mid \mathbf{x}(t_k)), \end{aligned} \tag{7.25}$$

which is a canonical discrete-time filtering problem (Särkkä, 2013)—provided that we introduce the notation $\mathbf{x}_k \triangleq \mathbf{x}(t_k)$.

The filtering distributions at times $t_1, t_2, \ldots$ can now be computed by starting from a prior distribution $p(\mathbf{x}(t_0))$ and by using the following Bayesian filter recursions (see Särkkä, 2013):

- *Initialization.* The recursion starts from the prior distribution $p(\mathbf{x}(t_0))$.
- *Prediction step.* The predictive distribution of the state $\mathbf{x}(t_k)$ at the time t_k, given the dynamic model, can be computed by the Chapman–Kolmogorov equation

$$\begin{aligned} &p(\mathbf{x}(t_k) \mid \mathbf{y}_1, \ldots, \mathbf{y}_{k-1}) = \\ &\qquad \int p(\mathbf{x}(t_k) \mid \mathbf{x}(t_{k-1}))\, p(\mathbf{x}(t_{k-1}) \mid \mathbf{y}_1, \ldots, \mathbf{y}_{k-1})\, \mathrm{d}\mathbf{x}(t_{k-1}). \end{aligned} \tag{7.26}$$

- *Update step.* Given the measurement $\mathbf{y}_k$ at time t_k the posterior distribution of the state $\mathbf{x}(t_k)$ can be computed by Bayes' rule

$$p(\mathbf{x}(t_k) \mid \mathbf{y}_1, \ldots, \mathbf{y}_k) = \frac{1}{Z_k}\, p(\mathbf{y}_k \mid \mathbf{x}(t_k))\, p(\mathbf{x}(t_k) \mid \mathbf{y}_1, \ldots, \mathbf{y}_{k-1}), \tag{7.27}$$

where the *normalization constant* Z_k is given by

$$Z_k = \int p(\mathbf{y}_k \mid \mathbf{x}(t_k))\, p(\mathbf{x}(t_k) \mid \mathbf{y}_1, \ldots, \mathbf{y}_{k-1})\, \mathrm{d}\mathbf{x}(t_k). \tag{7.28}$$

A linear Gaussian continuous/discrete-time filtering model has the general form

$$\begin{aligned} \mathrm{d}\mathbf{x} &= \mathbf{F}(t)\,\mathbf{x}\,\mathrm{d}t + \mathbf{L}(t)\,\mathrm{d}\boldsymbol{\beta}, \\ \mathbf{y}_k &= \mathbf{H}_k\,\mathbf{x}(t_k) + \mathbf{r}_k, \end{aligned} \tag{7.29}$$

where $\mathbf{r}_k \sim \mathrm{N}(\mathbf{0}, \mathbf{R}_k)$ and $\boldsymbol{\beta}$ is a Brownian motion with diffusion matrix $\mathbf{Q}$. This model is of the type which we already encountered in Example 7.2. From Sections 4.6 and 4.7 we now deduce that the corresponding discrete-time model has the form

$$\begin{aligned} \mathbf{x}(t_{k+1}) &= \mathbf{A}_k\,\mathbf{x}(t_k) + \mathbf{q}_k, \\ \mathbf{y}_k &= \mathbf{H}_k\,\mathbf{x}(t_k) + \mathbf{r}_k, \end{aligned} \tag{7.30}$$

where $\mathbf{q}_k \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_k)$ with $\mathbf{A}_k$ and $\boldsymbol{\Sigma}_k$ given by Equations (4.51) and (4.53) or in the linear time-invariant (LTI) case by (4.57) and (4.58). This model is thus a discrete-time version of the linear Gaussian model analogously to (7.25).

Assuming that $p(\mathbf{x}(0)) = \mathrm{N}(\mathbf{x}(t_0) \mid \mathbf{m}_0, \mathbf{P}_0)$, the corresponding filtering solution is now given by the following Kalman filter (Kalman, 1960; Särkkä, 2013).

- *Initialization.* The recursion is started from the prior mean $\mathbf{m}_0$ and covariance $\mathbf{P}_0$.
- *Prediction step.*

$$\begin{aligned} \mathbf{m}_k^- &= \mathbf{A}_{k-1}\,\mathbf{m}_{k-1}, \\ \mathbf{P}_k^- &= \mathbf{A}_{k-1}\,\mathbf{P}_{k-1}\,\mathbf{A}_{k-1}^\mathsf{T} + \boldsymbol{\Sigma}_{k-1}. \end{aligned} \tag{7.31}$$

- *Update step.*

$$\begin{aligned} \mathbf{v}_k &= \mathbf{y}_k - \mathbf{H}_k\,\mathbf{m}_k^-, \\ \mathbf{S}_k &= \mathbf{H}_k\,\mathbf{P}_k^-\,\mathbf{H}_k^\mathsf{T} + \mathbf{R}_k, \\ \mathbf{K}_k &= \mathbf{P}_k^-\,\mathbf{H}_k^\mathsf{T}\,\mathbf{S}_k^{-1}, \\ \mathbf{m}_k &= \mathbf{m}_k^- + \mathbf{K}_k\,\mathbf{v}_k, \\ \mathbf{P}_k &= \mathbf{P}_k^- - \mathbf{K}_k\,\mathbf{S}_k\,\mathbf{K}_k^\mathsf{T}. \end{aligned} \tag{7.32}$$

This procedure gives the following distributions:

$$\begin{aligned} p(\mathbf{x}(t_k) \mid \mathbf{y}_1, \ldots, \mathbf{y}_{k-1}) &= \mathrm{N}(\mathbf{x}(t_k) \mid \mathbf{m}_k^-, \mathbf{P}_k^-), \\ p(\mathbf{x}(t_k) \mid \mathbf{y}_1, \ldots, \mathbf{y}_k) &= \mathrm{N}(\mathbf{x}(t_k) \mid \mathbf{m}_k, \mathbf{P}_k), \\ p(\mathbf{y}_k \mid \mathbf{y}_1, \ldots, \mathbf{y}_{k-1}) &= \mathrm{N}(\mathbf{y}_k \mid \mathbf{H}_k\mathbf{m}_k^-, \mathbf{S}_k). \end{aligned} \tag{7.33}$$

Instead of first forming the equivalent discrete-time system it is possible to derive a Bayesian filter directly for the continuous-discrete filtering problem (7.19). The result is the following algorithm.

- *Prediction step.* Solve the predicted probability density at time of the measurement $p(\mathbf{x}(t_k) \mid \mathbf{y}_1, \ldots, \mathbf{y}_{k-1})$ by integrating the Fokker–Planck–Kolmogorov equation (see Theorem 4.1) from the filtering density at the previous measurement time step t_{k-1} to the current time t_k:

$$\begin{aligned}\frac{\partial p(\mathbf{x}, t)}{\partial t} &= -\sum_i \frac{\partial}{\partial x_i}[f_i(x, t)\, p(\mathbf{x}, t)] \\ &\quad + \frac{1}{2}\sum_{ij} \frac{\partial^2}{\partial x_i\, \partial x_j} \left\{[\mathbf{L}(\mathbf{x}, t)\, \mathbf{Q}\, \mathbf{L}^{\mathsf{T}}(\mathbf{x}, t)]_{ij}\; p(\mathbf{x}, t)\right\},\end{aligned} \tag{7.34}$$

  where we have denoted the filtering distribution as $p(\mathbf{x}, t) \triangleq p(\mathbf{x}(t) \mid \mathbf{y}_1, \ldots, \mathbf{y}_{k-1})$ and the initial condition is $p(\mathbf{x}, t_{k-1}) \triangleq p(\mathbf{x}(t_k) \mid \mathbf{y}_1, \ldots, \mathbf{y}_{k-1})$.

- *Update step.* Use Bayes' rule (7.27) for calculating the conditional distribution $p(\mathbf{x}(t_k) \mid \mathbf{y}_1, \ldots, \mathbf{y}_k)$, given the new measurement $\mathbf{y}_k$.

The corresponding continuous/discrete Kalman filter solution to the linear Gaussian model (7.29) is the following.

- *Prediction step.* The differential equations

$$\begin{aligned}\frac{\mathrm{d}\mathbf{m}(t)}{\mathrm{d}t} &= \mathbf{F}(t)\, \mathbf{m}(t), \\ \frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t} &= \mathbf{F}(t)\, \mathbf{P}(t) + \mathbf{P}(t)\, \mathbf{F}^T(t) + \mathbf{L}(t)\, \mathbf{Q}(t)\, \mathbf{L}^T(t),\end{aligned} \tag{7.35}$$

  are integrated from the initial conditions $\mathbf{m}(t_{k-1}) = \mathbf{m}_{k-1}$, $\mathbf{P}(t_{k-1}) = \mathbf{P}_{k-1}$ to time instance t_k. The predicted mean and covariance are given as $\mathbf{m}_k^- = \mathbf{m}(t_k)$ and $\mathbf{P}_k^- = \mathbf{P}(t_k)$, respectively.

- *Update step.* The update step is the same as the discrete-time Kalman filter update step given in Equations (7.32).

7.5 Continuous/discrete-time approximate non-linear filtering

One approach to approximate non-linear continuous/discrete filtering is to approximate the transition density using Itô–Taylor or stochastic Runge–Kutta approximations discussed in the previous chapters. For example, from the Euler–Maruyama discretization in the model (7.19) we get the approximate model

$$\begin{aligned}\mathbf{x}(t_{k+1}) &= \mathbf{x}(t_k) + \mathbf{f}(\mathbf{x}(t_k), t_k)\, \Delta t + \mathbf{L}(\mathbf{x}(t_k), t_k)\, \Delta \boldsymbol{\beta}_k, \\ \mathbf{y}_k &\sim p(\mathbf{y}_k \mid \mathbf{x}(t_k)),\end{aligned} \tag{7.36}$$

which corresponds to a discrete-time model (7.25) with

$$p(\mathbf{x}(t_{k+1}) \mid \mathbf{x}(t_k)) = \mathrm{N}(\mathbf{x}(t_{k+1}) \mid \mathbf{x}(t_k) + \mathbf{f}(\mathbf{x}(t_k), t_k)\,\Delta t, \mathbf{L}(\mathbf{x}(t_k), t_k)\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}(\mathbf{x}(t_k), t_k)\,\Delta t). \quad (7.37)$$

Weak Ito–Taylor series and Runge–Kutta approximations similarly correspond to transition density approximations with either Gaussian or binomial/trinomial noises.

Another classical approach (see, *e.g.*, Jazwinski, 1970) is to replace the Fokker–Planck–Kolmogorov equation solution in Equation (7.34) with linearization and sigma-point approximations of SDEs presented in Section (5.2). This approach was also more recently studied in Särkkä (2006); Särkkä and Sarmavuori (2013) and a comparison to the discretization approach above was reported in Särkkä and Solin (2012).

7.6 Bayesian smoothing

Solving a Bayesian smoothing problem means computation of the distributions

$$p(\mathbf{x}(t) \mid \mathbf{y}_1, \ldots, \mathbf{y}_T), \qquad t \in [t_0, t_T], \quad (7.38)$$

or in continuous case

$$p(\mathbf{x}(t) \mid \mathcal{Y}_T), \qquad t \in [t_0, T], \quad (7.39)$$

that is, computation of the posterior distributions of the states within the range of all measurements. For example, in the car tracking problem this corresponds to determination of the position and velocity of the car during some past points of time, but still by using information from all the measurements obtained so far. That is, the smoothing problem is roughly equivalent to batch estimation of the state although the algorithms is formulated a bit differently to gain better computational scaling.

The solutions to these problems are classical, and reviews of the classical and more recent methods can be found, for example, in the references Särkkä (2013) and Särkkä and Sarmavuori (2013).

7.7 Parameter estimation

Parameter estimation in the continuous/discrete-time case (we only consider that here) is considered with models of the form

$$\begin{aligned} \mathrm{d}\mathbf{x} &= \mathbf{f}(\mathbf{x}, t; \boldsymbol{\theta})\,\mathrm{d}t + \mathbf{L}(\mathbf{x}, t; \boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\beta}(t), \\ \mathbf{y}_k &\sim p(\mathbf{y}_k \mid \mathbf{x}(t_k); \boldsymbol{\theta}), \end{aligned} \quad (7.40)$$

where $\boldsymbol{\theta}$ is now a vector of unknown parameters of the system. Luckily it turns out that provided that we can solve the filtering problem, we can also solve the corresponding parameter estimation problem.

The basic idea is to notice that the normalization constant Z_k in Equation (7.28), which is a by-product of the filtering equations, is actually

$$\begin{aligned} Z_k(\boldsymbol{\theta}) &= \int p(\mathbf{y}_k \mid \mathbf{x}(t_k); \boldsymbol{\theta})\, p(\mathbf{x}(t_k) \mid \mathbf{y}_1, \ldots, \mathbf{y}_{k-1}; \boldsymbol{\theta})\, \mathrm{d}\mathbf{x}(t_k) \\ &= p(\mathbf{y}_k \mid \mathbf{y}_1, \ldots, \mathbf{y}_{k-1}; \boldsymbol{\theta}). \end{aligned} \tag{7.41}$$

Thus we can express the marginal likelihood simply as

$$p(\mathbf{y}_1, \ldots, \mathbf{y}_k \mid \boldsymbol{\theta}) = \prod_k p(\mathbf{y}_k \mid \mathbf{y}_1, \ldots, \mathbf{y}_{k-1}; \boldsymbol{\theta}) = \prod_k Z_k(\boldsymbol{\theta}). \tag{7.42}$$

Given the marginal likelihood the posterior distribution is given as

$$p(\boldsymbol{\theta} \mid \mathbf{y}_1, \ldots, \mathbf{y}_k) \propto p(\mathbf{y}_1, \ldots, \mathbf{y}_k \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}), \tag{7.43}$$

which can further be optimized or sampled using methods like Markov chain Monte Carlo (MCMC). For more information the reader is referred to Särkkä (2013) and Särkkä et al. (2014).

Chapter 8

Further topics

8.1 Martingale properties and generators of SDEs

In this section we discuss martingale properties and generators of Itô processes which are important concepts in theoretical analysis of SDEs. The definition of a martingale is the following.

Definition 8.1 (Martingale)**.** *A stochastic process* $\mathbf{x}(t)$ *with bounded expectation* $\mathrm{E}[\mathbf{x}(t)] < \infty$ *is called a* martingale *if*

$$\mathrm{E}[\mathbf{x}(t) \mid \mathcal{X}_s] = \mathbf{x}(s), \qquad \text{for all } t \geq s. \tag{8.1}$$

It turns out that all Itô integrals are martingales, and this follows from the fact that Brownian motion is a martingale as well. However, solutions to SDEs are martingales only if the drift $\mathbf{f}(\mathbf{x}, t) = 0$. For more information on martingales and their role in stochastic calculus reader is referred to Øksendal (2003) and Karatzas and Shreve (1991).

Yet another useful concept in the theory of Itô processes and more general stochastic processes is the (infinitesimal) generator which in the case of (time-invariant) Itô diffusions is the following.

Definition 8.2 (Generator)**.** *The generator of a time-invariant stochastic process is defined as*

$$\mathcal{A}\,\phi(\mathbf{x}) = \lim_{t \downarrow 0} \frac{\mathrm{E}[\phi(x(t))] - \phi(x(0))}{t}. \tag{8.2}$$

For a time-invariant Itô process defined as the solution to the SDE

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x})\,\mathrm{d}t + \mathbf{L}(\mathbf{x})\,\mathrm{d}\boldsymbol{\beta}, \tag{8.3}$$

the generator is given as

$$\mathcal{A}(\bullet) = \sum_i \mathrm{E}\left[\frac{\partial(\bullet)}{\partial x_i}\, f_i(\mathbf{x})\right] + \frac{1}{2}\sum_{ij} \mathrm{E}\left[\left(\frac{\partial^2(\bullet)}{\partial x_i\, \partial x_j}\right) [\mathbf{L}(\mathbf{x})\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}(\mathbf{x})]_{ij}\right] \tag{8.4}$$

which we in fact already encountered in Equation (4.4)*.*

8.2 Girsanov theorem

The purpose of this section is to present an intuitive derivation of the Girsanov theorem, which is a very useful theorem, but in its general form, slightly abstract. The derivation should only be considered to be an attempt to reveal the intuition behind the theorems, and not be considered an actual proof of the theorem. The derivation is based on considering formal probability densities of paths of Itô processes, which is intuitive, but not really the mathematically correct way to go. To be rigorous, we should not attempt to consider probability densities of the paths at all, but instead consider the probability measures of the processes (*cf.* Øksendal, 2003).

The Girsanov theorem (Theorem 8.3) is due to Girsanov (1960), and in addition to the original article, its proper derivation can be found, for example, in Karatzas and Shreve (1991) (see also Øksendal, 2003). The derivation of Theorem 8.1 from the Girsanov theorem can be found in Särkkä and Sottinen (2008). Here we proceed backwards from Theorem 8.1 to Theorem 8.3.

Let's denote the whole path of the Itô process $\mathbf{x}(t)$ on a time interval $[0, t]$ as follows:

$$\mathcal{X}_t = \{\mathbf{x}(\tau) \ : \ 0 \leq \tau \leq t\}. \tag{8.5}$$

Let $\mathbf{x}(t)$ be the solution to

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\,\mathrm{d}t + \mathrm{d}\boldsymbol{\beta}. \tag{8.6}$$

Here we have set $\mathbf{L}(\mathbf{x}, t) = \mathbf{I}$ for notational simplicity. In fact, Girsanov theorem can be used for general time varying $\mathbf{L}(t)$ provided that $\mathbf{L}(t)$ is invertible for each t. This invertibility requirement can also be relaxed in some situations (*cf.* Särkkä and Sottinen, 2008).

For any finite N, the joint probability density $p(\mathbf{x}(t_1), \ldots, \mathbf{x}(t_N))$ exists (provided that certain technical conditions are met) for an arbitrary finite collection of times $t_1, \ldots, t_N$. We will now formally define the probability density of the whole path as

$$p(\mathcal{X}_t) = \lim_{N \to \infty} p(\mathbf{x}(t_1), \ldots, \mathbf{x}(t_N)), \tag{8.7}$$

where the times $t_1, \ldots, t_N$ need to selected such that they become dense in the limit. In fact this density is not normalizable, but we can still define the density though the ratio between the joint probability density of $\mathbf{x}$ and another process $\mathbf{y}$:

$$\frac{p(\mathcal{X}_t)}{p(\mathcal{Y}_t)} = \lim_{N \to \infty} \frac{p(\mathbf{x}(t_1), \ldots, \mathbf{x}(t_N))}{p(\mathbf{y}(t_1), \ldots, \mathbf{y}(t_N))}. \tag{8.8}$$

This is a finite quantity with a suitable choice of $\mathbf{y}$. We can also denote the expectation of a functional $h(\mathcal{X}_t)$ of the path as follows:

$$\mathrm{E}[h(\mathcal{X}_t)] = \int h(\mathcal{X}_t)\, p(\mathcal{X}_t)\, \mathrm{d}\mathcal{X}_t. \tag{8.9}$$

In physics this kind of integrals are called path integrals (Chaichian and Demichev, 2001a,b). Note that this notation is purely formal, because the density $p(\mathcal{X}_t)$ is actually an infinite quantity. However, the expectation is indeed finite. Let's now compute the ratio of probability densities for a pair of processes.

Theorem 8.1 (Likelihood ratio of Itô processes)**.** *Consider the Itô processes*

$$\begin{aligned} \mathrm{d}\mathbf{x} &= \mathbf{f}(\mathbf{x}, t)\,\mathrm{d}t + \mathrm{d}\boldsymbol{\beta}, \qquad \mathbf{x}(0) = \mathbf{x}_0, \\ \mathrm{d}\mathbf{y} &= \mathbf{g}(\mathbf{y}, t)\,\mathrm{d}t + \mathrm{d}\boldsymbol{\beta}, \qquad \mathbf{y}(0) = \mathbf{x}_0, \end{aligned} \tag{8.10}$$

where the Brownian motion $\boldsymbol{\beta}(t)$ has a non-singular diffusion matrix $\mathbf{Q}$. Then the ratio of the probability laws of $\mathcal{X}_t$ and $\mathcal{Y}_t$ is given as

$$\frac{p(\mathcal{X}_t)}{p(\mathcal{Y}_t)} = Z(t), \tag{8.11}$$

where

$$Z(t) = \exp\left(-\frac{1}{2}\int_0^t [\mathbf{f}(\mathbf{y},\tau) - \mathbf{g}(\mathbf{y},\tau)]^\mathsf{T}\,\mathbf{Q}^{-1}\,[\mathbf{f}(\mathbf{y},\tau) - \mathbf{g}(\mathbf{y},\tau)]\,\mathrm{d}\tau \right. \\ \left. + \int_0^t [\mathbf{f}(\mathbf{y},\tau) - \mathbf{g}(\mathbf{y},\tau)]^\mathsf{T}\,\mathbf{Q}^{-1}\,\mathrm{d}\boldsymbol{\beta}(\tau) \right) \tag{8.12}$$

in the sense that for an arbitrary functional $h(\bullet)$ of the path from 0 to t we have

$$\mathrm{E}[h(\mathcal{X}_t)] = \mathrm{E}[Z(t)\,h(\mathcal{Y}_t)], \tag{8.13}$$

where the expectation is over the randomness induced by the Brownian motion. Furthermore, under the probability measure defined through the transformed probability density

$$\tilde{p}(\mathcal{X}_t) = Z(t)\,p(\mathcal{X}_t) \tag{8.14}$$

the process

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta} - \int_0^t [\mathbf{f}(\mathbf{y},\tau) - \mathbf{g}(\mathbf{y},\tau)]\,\mathrm{d}\tau \tag{8.15}$$

is a Brownian motion with diffusion matrix $\mathbf{Q}$.

Derivation. Let's discretize the time interval $[0, t]$ into N time steps $0 = t_0 < t_1 < \ldots < t_N = t$, where $t_{i+1} - t_i = \Delta t$, and let's denote $\mathbf{x}_i = \mathbf{x}(t_i)$ and $\mathbf{y}_i = \mathbf{y}(t_i)$. When Δt is small, we have

$$\begin{aligned} p(\mathbf{x}_{i+1} \mid \mathbf{x}_i) &= \mathrm{N}(\mathbf{x}_{i+1} \mid \mathbf{x}_i + \mathbf{f}(\mathbf{x}_i, t)\,\Delta t, \mathbf{Q}\,\Delta t), \\ q(\mathbf{y}_{i+1} \mid \mathbf{y}_i) &= \mathrm{N}(\mathbf{y}_{i+1} \mid \mathbf{y}_i + \mathbf{g}(\mathbf{y}_i, t)\,\Delta t, \mathbf{Q}\,\Delta t). \end{aligned} \tag{8.16}$$

The joint density p of $\mathbf{x}_1, \ldots, \mathbf{x}_N$ can then be written in the form

$$\begin{aligned} p(\mathbf{x}_1, \ldots, \mathbf{x}_N) &= \prod_i \mathrm{N}(\mathbf{x}_{i+1} \mid \mathbf{x}_i + \mathbf{f}(\mathbf{x}_i, t)\,\Delta t, \mathbf{Q}\,\Delta t) \\ &= |2\pi\,\mathbf{Q}\,\mathrm{d}t|^{-N/2} \exp\Bigg(-\frac{1}{2}\sum_i (\mathbf{x}_{i+1} - \mathbf{x}_i)^{\mathsf{T}} (\mathbf{Q}\,\Delta t)^{-1} (\mathbf{x}_{i+1} - \mathbf{x}_i) \\ &\qquad -\frac{1}{2}\sum_i \mathbf{f}^{\mathsf{T}}(\mathbf{x}_i, t_i)\,\mathbf{Q}^{-1}\mathbf{f}(\mathbf{x}_i, t_i)\,\Delta t + \sum_i \mathbf{f}^{\mathsf{T}}(\mathbf{x}_i, t_i)\,\mathbf{Q}^{-1}(\mathbf{x}_{i+1} - \mathbf{x}_i)\Bigg) \end{aligned} \tag{8.17}$$

For the joint density q of $\mathbf{y}_1, \ldots, \mathbf{y}_N$ we similarly get

$$\begin{aligned} q(\mathbf{y}_1, \ldots, \mathbf{y}_N) &= \prod_i \mathrm{N}(\mathbf{y}_{i+1} \mid \mathbf{y}_i + \mathbf{g}(\mathbf{y}_i, t)\,\Delta t, \mathbf{Q}\,\Delta t) \\ &= |2\pi\,\mathbf{Q}\,\mathrm{d}t|^{-N/2} \exp\Bigg(-\frac{1}{2}\sum_i (\mathbf{y}_{i+1} - \mathbf{y}_i)^{\mathsf{T}} (\mathbf{Q}\,\Delta t)^{-1} (\mathbf{y}_{i+1} - \mathbf{y}_i) \\ &\qquad -\frac{1}{2}\sum_i \mathbf{g}^{\mathsf{T}}(\mathbf{y}_i, t_i)\,\mathbf{Q}^{-1}\mathbf{g}(\mathbf{y}_i, t_i)\,\Delta t + \sum_i \mathbf{g}^{\mathsf{T}}(\mathbf{y}_i, t_i)\,\mathbf{Q}^{-1}(\mathbf{y}_{i+1} - \mathbf{y}_i)\Bigg) \end{aligned} \tag{8.18}$$

For any function h_N we have

$$\begin{aligned} &\int h_N(\mathbf{x}_1, \ldots, \mathbf{x}_N)\, p(\mathbf{x}_1, \ldots, \mathbf{x}_N)\, \mathrm{d}(\mathbf{x}_1, \ldots, \mathbf{x}_N) \\ &= \int h_N(\mathbf{x}_1, \ldots, \mathbf{x}_N)\, \frac{p(\mathbf{x}_1, \ldots, \mathbf{x}_N)}{q(\mathbf{x}_1, \ldots, \mathbf{x}_N)}\, q(\mathbf{x}_1, \ldots, \mathbf{x}_N)\, \mathrm{d}(\mathbf{x}_1, \ldots, \mathbf{x}_N) \\ &= \int h_N(\mathbf{y}_1, \ldots, \mathbf{y}_N)\, \frac{p(\mathbf{y}_1, \ldots, \mathbf{y}_N)}{q(\mathbf{y}_1, \ldots, \mathbf{y}_N)}\, q(\mathbf{y}_1, \ldots, \mathbf{y}_N)\, \mathrm{d}(\mathbf{y}_1, \ldots, \mathbf{y}_N). \end{aligned} \tag{8.19}$$

Thus we still only need to consider the following:

$$\begin{aligned} &\frac{p(\mathbf{y}_1, \ldots, \mathbf{y}_N)}{q(\mathbf{y}_1, \ldots, \mathbf{y}_N)} \\ &= \exp\Bigg(-\frac{1}{2}\sum_i \mathbf{f}^{\mathsf{T}}(\mathbf{y}_i, t_i)\,\mathbf{Q}^{-1}\mathbf{f}(\mathbf{y}_i, t_i)\,\Delta t + \sum_i \mathbf{f}^{\mathsf{T}}(\mathbf{y}_i, t_i)\,\mathbf{Q}^{-1}(\mathbf{y}_{i+1} - \mathbf{y}_i) \\ &\qquad + \frac{1}{2}\sum_i \mathbf{g}^{\mathsf{T}}(\mathbf{y}_i, t_i)\,\mathbf{Q}^{-1}\mathbf{g}(\mathbf{y}_i, t_i)\,\Delta t - \sum_i \mathbf{g}^{\mathsf{T}}(\mathbf{y}_i, t_i)\,\mathbf{Q}^{-1}(\mathbf{y}_{i+1} - \mathbf{y}_i)\Bigg). \end{aligned} \tag{8.20}$$

In the limit $N \to \infty$ the exponential becomes

$$
\begin{aligned}
&-\frac{1}{2}\sum_i \mathbf{f}^\mathsf{T}(\mathbf{y}_i,t_i)\,\mathbf{Q}^{-1}\mathbf{f}(\mathbf{y}_i,t_i)\,\Delta t + \sum_i \mathbf{f}^\mathsf{T}(\mathbf{y}_i,t_i)\,\mathbf{Q}^{-1}(\mathbf{y}_{i+1}-\mathbf{y}_i)\\
&\quad+\frac{1}{2}\sum_i \mathbf{g}^\mathsf{T}(\mathbf{y}_i,t_i)\,\mathbf{Q}^{-1}\mathbf{g}(\mathbf{y}_i,t_i)\,\Delta t - \sum_i \mathbf{g}^\mathsf{T}(\mathbf{y}_i,t_i)\,\mathbf{Q}^{-1}(\mathbf{y}_{i+1}-\mathbf{y}_i)\\
&\to -\frac{1}{2}\int_0^t \mathbf{f}^\mathsf{T}(\mathbf{y},\tau)\,\mathbf{Q}^{-1}\mathbf{f}(\mathbf{y},\tau)\,\mathrm{d}\tau + \int_0^t \mathbf{f}^\mathsf{T}(\mathbf{y},\tau)\,\mathbf{Q}^{-1}\,\mathrm{d}\mathbf{y}\\
&\quad+\frac{1}{2}\int_0^t \mathbf{g}^\mathsf{T}(\mathbf{y},\tau)\,\mathbf{Q}^{-1}\,\mathbf{g}(\mathbf{y},\tau)\,\mathrm{d}\tau - \int_0^t \mathbf{g}^\mathsf{T}(\mathbf{y},\tau)\,\mathbf{Q}^{-1}\,\mathrm{d}\mathbf{y}\\
&= -\frac{1}{2}\int_0^t [\mathbf{f}(\mathbf{y},\tau)-\mathbf{g}(\mathbf{y},\tau)]^\mathsf{T}\,\mathbf{Q}^{-1}\,[\mathbf{f}(\mathbf{y},\tau)-\mathbf{g}(\mathbf{y},\tau)]\,\mathrm{d}\tau\\
&\quad+\int_0^t [\mathbf{f}(\mathbf{y},\tau)-\mathbf{g}(\mathbf{y},\tau)]^\mathsf{T}\,\mathbf{Q}^{-1}\,\mathrm{d}\boldsymbol{\beta},
\end{aligned}
\tag{8.21}
$$

where we have substituted $\mathrm{d}\mathbf{y} = \mathbf{g}(\mathbf{y},t)\,\mathrm{d}t + \mathrm{d}\boldsymbol{\beta}$. Thus this gives the expression for $Z(t)$. We can now solve the Brownian motion $\boldsymbol{\beta}$ from the first SDE as

$$
\boldsymbol{\beta}(t) = \mathbf{x}(t) - \mathbf{x}_0 - \int_0^t \mathbf{f}(\mathbf{x},\tau)\,\mathrm{d}\tau. \tag{8.22}
$$

The expectation of an arbitrary functional h of the Brownian motion can now be expressed as

$$
\begin{aligned}
\mathrm{E}[h(\mathcal{B}_t)] &= \mathrm{E}\left[h\left(\left\{\mathbf{x}(s)-\mathbf{x}_0-\int_0^s \mathbf{f}(\mathbf{x},\tau)\,\mathrm{d}\tau \;:\; 0\le s\le t\right\}\right)\right]\\
&= \mathrm{E}\left[Z(t)\,h\left(\left\{\mathbf{y}(s)-\mathbf{x}_0-\int_0^s \mathbf{f}(\mathbf{y},\tau)\,\mathrm{d}\tau \;:\; 0\le s\le t\right\}\right)\right]\\
&= \mathrm{E}\left[Z(t)\,h\left(\left\{\int_0^s \mathbf{g}(\mathbf{y},\tau)+\boldsymbol{\beta}(t)-\int_0^s \mathbf{f}(\mathbf{y},\tau)\,\mathrm{d}\tau \;:\; 0\le s\le t\right\}\right)\right]\\
&= \mathrm{E}\left[Z(t)\,h\left(\left\{\boldsymbol{\beta}(t)-\int_0^s [\mathbf{f}(\mathbf{y},\tau)-\mathbf{g}(\mathbf{y},\tau)]\;\mathrm{d}\tau \;:\; 0\le s\le t\right\}\right)\right],
\end{aligned}
\tag{8.23}
$$

which implies that $\boldsymbol{\beta}(t) - \int_0^s [\mathbf{f}(\mathbf{y},\tau)-\mathbf{g}(\mathbf{y},\tau)]\,\mathrm{d}\tau$ has the statistics of Brownian motion provided that we scale the probability density with $Z(t)$. □

Remark 8.1. *We need to have*

$$
\begin{aligned}
\mathrm{E}\left[\exp\left(\int_0^t \mathbf{f}(\mathbf{y},\tau)^\mathsf{T}\,\mathbf{Q}^{-1}\,\mathbf{f}(\mathbf{y},\tau)\,\mathrm{d}\tau\right)\right] &< \infty,\\
\mathrm{E}\left[\exp\left(\int_0^t \mathbf{g}(\mathbf{y},\tau)^\mathsf{T}\,\mathbf{Q}^{-1}\,\mathbf{g}(\mathbf{y},\tau)\,\mathrm{d}\tau\right)\right] &< \infty,
\end{aligned}
\tag{8.24}
$$

because otherwise $Z(t)$ will be zero.

Let's now write a slightly more abstract version of the above theorem which is roughly equivalent to the actual Girsanov theorem in the form that it is usually found in stochastics literature.

Theorem 8.2 (Girsanov I)**.** *Let $\boldsymbol{\theta}(t)$ be a process which is driven by a standard Brownian motion $\boldsymbol{\beta}(t)$ such that*

$$\mathrm{E}\left[\int_0^t \boldsymbol{\theta}^{\mathsf{T}}(\tau)\,\boldsymbol{\theta}(\tau)\,\mathrm{d}\tau\right] < \infty, \tag{8.25}$$

then under the measure defined by the formal probability density

$$\tilde{p}(\Theta_t) = Z(t)\; p(\Theta_t), \tag{8.26}$$

where $\Theta_t = \{\boldsymbol{\theta}(\tau)\ :\ 0 \le \tau \le t\}$, and

$$Z(t) = \exp\left(\int_0^t \boldsymbol{\theta}^{\mathsf{T}}(\tau)\,\mathrm{d}\boldsymbol{\beta} - \frac{1}{2}\int_0^t \boldsymbol{\theta}^{\mathsf{T}}(\tau)\,\boldsymbol{\theta}(\tau)\,\mathrm{d}\tau\right), \tag{8.27}$$

the following process is a standard Brownian motion:

$$\tilde{\boldsymbol{\beta}}(t) = \boldsymbol{\beta}(t) - \int_0^t \boldsymbol{\theta}(\tau)\,\mathrm{d}\tau. \tag{8.28}$$

Derivation. Select $\boldsymbol{\theta}(t) = \mathbf{f}(\mathbf{y}, t) - \mathbf{g}(\mathbf{y}, t)$ and $\mathbf{Q} = \mathbf{I}$ in the previous theorem. □

In fact, the above derivation does not yet guarantee that any selected $\boldsymbol{\theta}(t)$ can be constructed like this. But still, it reveals the link between the likelihood ratio and the Girsanov theorem. Despite the limited derivation the above theorem is generally true. The detailed technical conditions can be found in the original article of Girsanov (1960).

However, the above theorem is still in the heuristic notation in terms of the formal probability densities of paths. In the proper formulation of the theorem $\boldsymbol{\theta}$ being "driven" by Brownian motion actually means that the process $\boldsymbol{\theta}$ is *adapted* to the Brownian motion. To be more explicit in notation, it is also common to write down the event space element $\omega \in \Omega$ as the argument of $\boldsymbol{\beta}(\omega, t)$. The processes $\boldsymbol{\theta}(\omega, t)$ and $Z(\omega, t)$ should then be functions of the event space element as well. In fact, $\boldsymbol{\theta}(\omega, t)$ should be non-anticipative (not looking into the future) functional of Brownian motion, that is, adapted to the natural filtration $\mathcal{F}_t$ of the Brownian motion. Furthermore, the ratio of probability densities is in fact the Radon–Nikodym derivative of the measure $\tilde{\mathbb{P}}(\omega)$ with respect to the other measure $\mathbb{P}(\omega)$. With these notations the Girsanov theorem looks like the following which is roughly the format found in stochastic books.

Theorem 8.3 (Girsanov II)**.** *Let $\boldsymbol{\beta}(\omega, t)$ be a standard n-dimensional Brownian motion under the probability measure $\mathbb{P}$. Let $\boldsymbol{\theta} : \Omega \times \mathbb{R}_+ \mapsto \mathbb{R}^n$ be an adapted process such that the process Z defined as*

$$Z(\omega, t) = \exp\left\{\int_0^t \boldsymbol{\theta}^\mathsf{T}(\omega, t)\,\mathrm{d}\boldsymbol{\beta}(\omega, t) - \frac{1}{2}\int_0^t \boldsymbol{\theta}^\mathsf{T}(\omega, t)\,\boldsymbol{\theta}(\omega, t)\,\mathrm{d}t\right\}, \tag{8.29}$$

satisfies $\mathrm{E}[Z(\omega, t)] = 1$*. Then the process*

$$\tilde{\boldsymbol{\beta}}(\omega, t) = \boldsymbol{\beta}(\omega, t) - \int_0^t \boldsymbol{\theta}(\omega, \tau)\,\mathrm{d}\tau \tag{8.30}$$

is a standard Brownian motion under the probability measure $\tilde{\mathbb{P}}$ defined via the relation

$$\mathrm{E}\left[\frac{\mathrm{d}\tilde{\mathbb{P}}}{\mathrm{d}\mathbb{P}}(\omega)\,\middle|\,\mathcal{F}_t\right] = Z(\omega, t), \tag{8.31}$$

where $\mathcal{F}_t$ is the natural filtration of the Brownian motion $\boldsymbol{\beta}(\omega, t)$.

8.3 Applications of the Girsanov theorem

The Girsanov theorem can be used for eliminating the drift functions and for finding weak solutions to SDEs by changing the measure suitably (see, *e.g.*, Øksendal, 2003; Särkkä, 2006). The basic idea in drift removal is to define $\boldsymbol{\theta}(t)$ in terms of the drift function suitably such that in the transformed SDE the drift cancels out. Construction of weak solutions is based on the result the process $\tilde{\boldsymbol{\beta}}(t)$ is a Brownian motion under the transformed measure. We can select $\boldsymbol{\theta}(t)$ such that there is another easily constructed process which then serves as the corresponding $\tilde{\mathbf{x}}(t)$ which solves the SDE driven by this new Brownian motion.

The Girsanov theorem is also important in stochastic filtering theory (see Ch. 7). The theorem can be used as the starting point of deriving the so-called Kallianpur–Striebel formula (Bayes' rule in continuous time). From this we can derive the whole stochastic filtering theory. The formula can also be used to form Monte Carlo (particle) methods to approximate filtering solutions. For details, see Crisan and Rozovskii (2011). In so called continuous-discrete filtering (continuous-time dynamics, discrete-time measurements) the theorem has turned out to be useful in constructing importance sampling and exact sampling methods for conditioned SDEs (Beskos et al., 2006; Särkkä and Sottinen, 2008).

8.4 Feynman–Kac formulae and parabolic PDEs

Feynman–Kac formula (see, *e.g.*, Øksendal, 2003) gives a link between solutions of parabolic partial differential equations (PDEs) and certain expected values of SDE solutions. In this section we shall present the general idea by deriving the

scalar Feynman–Kac equation. The multidimensional version could be obtained in an analogous way.

Let's start by considering the following PDE for function $u(x,t)$:

$$\begin{aligned} \frac{\partial u}{\partial t} + f(x)\,\frac{\partial u}{\partial x} + \frac{1}{2}\sigma^2(x)\frac{\partial^2 u}{\partial x^2} &= 0, \\ u(x,T) &= \Psi(x), \end{aligned} \tag{8.32}$$

where $f(x)$, $\sigma(x)$ and $\Psi(x)$ are some given functions and T is a fixed time instant. Let's define a process $x(t)$ on the interval $[t', T]$ as follows:

$$\mathrm{d}x = f(x)\,\mathrm{d}t + \sigma(x)\,\mathrm{d}\beta, \qquad x(t') = x', \tag{8.33}$$

that is, the process starts from a given x' at time t'. Let's now use the Itô formula for $u(x,t)$, and recall that it solves the PDE (8.32) which gives:

$$\begin{aligned} \mathrm{d}u &= \frac{\partial u}{\partial t}\,\mathrm{d}t + \frac{\partial u}{\partial x}\,\mathrm{d}x + \frac{1}{2}\frac{\partial^2 u}{\partial x^2}\,\mathrm{d}x^2 \\ &= \frac{\partial u}{\partial t}\,\mathrm{d}t + \frac{\partial u}{\partial x}\,f(x)\,\mathrm{d}t + \frac{\partial u}{\partial x}\,\sigma(x)\,\mathrm{d}\beta + \frac{1}{2}\frac{\partial^2 u}{\partial x^2}\,\sigma^2(x)\,\mathrm{d}t \\ &= \underbrace{\left[\frac{\partial u}{\partial t} + f(x)\,\frac{\partial u}{\partial x} + \frac{1}{2}\sigma^2(x)\,\frac{\partial^2 u}{\partial x^2}\right]}_{=0}\mathrm{d}t + \frac{\partial u}{\partial x}\,\sigma(x)\,\mathrm{d}\beta \\ &= \frac{\partial u}{\partial x}\,\sigma(x)\,\mathrm{d}\beta. \end{aligned} \tag{8.34}$$

Integrating from t' to T now gives

$$u(x(T),T) - u(x(t'),t') = \int_{t'}^{T} \frac{\partial u}{\partial x}\,\sigma(x)\,\mathrm{d}\beta, \tag{8.35}$$

and by substituting the initial and terminal terms we get:

$$\Psi(x(T)) - u(x',t') = \int_{t'}^{T} \frac{\partial u}{\partial x}\,\sigma(x)\,\mathrm{d}\beta. \tag{8.36}$$

We can now take expectations from both sides and recall that the expectation of any Itô integral is zero. Thus after rearranging we get

$$u(x',t') = \mathrm{E}[\Psi(x(T))]. \tag{8.37}$$

This means that we can solve the value of $u(x',t')$ for arbitrary x' and t' by starting the process in Equation (8.33) from x' and time t' and letting it run until time T. The solution is then the expected value of $\Psi(x(T))$ over the process realizations.

The above idea can also be generalized to equations of the form

$$\begin{aligned} \frac{\partial u}{\partial t} + f(x)\,\frac{\partial u}{\partial x} + \frac{1}{2}\sigma^2(x)\frac{\partial^2 u}{\partial x^2} - r\,u &= 0, \\ u(x,T) &= \Psi(x), \end{aligned} \tag{8.38}$$

where r is a positive constant. The corresponding SDE will be the same, but we need to apply the Itô formula to $\exp(-r\,t)\,u(x,t)$ instead of $u(x,t)$. The resulting *Feynman–Kac equation* is

$$u(x',t') = \exp(-r\,(T-t'))\;\mathrm{E}[\Psi(x(T))]. \tag{8.39}$$

We can generalize even more and allow r to depend on x, include additional constant term to the PDE and so on. Anyway, the idea remains the same.

8.5 Solving boundary value problems with Feynman–Kac

The Feynman–Kac equation can also be used for computing solutions to boundary value problems which do not include time variables at all (see, *e.g.*, Øksendal, 2003). Also in this section we only consider the scalar case, but analogous derivation works for the multi-dimensional case as well. Furthermore, proper derivation of the results in this section would need us to define the concept of random stopping time, which we have not done and thus in this sense the derivation is quite heuristic.

Let's consider the following boundary value problem for a function $u(x)$ defined on some finite domain Ω with boundary $\partial\Omega$:

$$\begin{aligned} f(x)\,\frac{\partial u}{\partial x} + \frac{1}{2}\sigma^2(x)\frac{\partial^2 u}{\partial x^2} &= 0 \\ u(x) &= \Psi(x), \quad x \in \partial\Omega\,. \end{aligned} \tag{8.40}$$

Again, let's define a process $x(t)$ in the same way as in Equation (8.33). Further, the application of the Itô formula to $u(x)$ gives

$$\begin{aligned} \mathrm{d}u &= \frac{\partial u}{\partial x}\,\mathrm{d}x + \frac{1}{2}\frac{\partial^2 u}{\partial x^2}\,\mathrm{d}x^2 \\ &= \frac{\partial u}{\partial x}\,f(x)\,\mathrm{d}t + \frac{\partial u}{\partial x}\,\sigma(x)\,\mathrm{d}\beta + \frac{1}{2}\frac{\partial^2 u}{\partial x^2}\,\sigma^2(x)\,\mathrm{d}t \\ &= \underbrace{\left[f(x)\,\frac{\partial u}{\partial x} + \frac{1}{2}\sigma^2(x)\,\frac{\partial^2 u}{\partial x^2}\right]}_{=0}\mathrm{d}t + \frac{\partial u}{\partial x}\,\sigma(x)\,\mathrm{d}\beta \\ &= \frac{\partial u}{\partial x}\,\sigma(x)\,\mathrm{d}\beta. \end{aligned} \tag{8.41}$$

Let T_e be the first exit time of the process $x(t)$ from the domain Ω. Integration from t' to T_e gives

$$u(x(T_e)) - u(x(t')) = \int_{t'}^{T_e} \frac{\partial u}{\partial x}\,\sigma(x)\,\mathrm{d}\beta. \tag{8.42}$$

But the value of $u(x)$ on the boundary is $\Psi(x)$ and $x(t') = x'$ thus we have

$$\Psi(x(T_e)) - u(x') = \int_{t'}^{T_e} \frac{\partial u}{\partial x}\, \sigma(x)\, \mathrm{d}\beta. \tag{8.43}$$

Taking expectation and rearranging then gives

$$u(x') = \mathrm{E}[\Psi(x(T_e))]. \tag{8.44}$$

That is, the value $u(x')$ with arbitrary x' can be obtained by starting the process $x(t)$ from $x(t') = x'$ in Equation (8.33) at arbitrary time t' and computing the expectation of $\Psi(x(T_e))$ over the first exit points of the process $x(t)$ from the domain Ω.

Again, we can generalize the derivation to equations of the form

$$\begin{aligned} f(x)\,\frac{\partial u}{\partial x} + \frac{1}{2}\sigma^2(x)\frac{\partial^2 u}{\partial x^2} - r\,u &= 0, \\ u(x) &= \Psi(x), \quad x \in \partial\Omega, \end{aligned} \tag{8.45}$$

which gives

$$u(x') = \exp(-r\,(T - t'))\ \mathrm{E}[\Psi(x(T_e))]. \tag{8.46}$$

8.6 Series expansions of Brownian motion

If we fix the time interval $[t_0, t_1]$ then on that interval standard Brownian motion has a Karhunen–Loeve series expansion of the form (see, *e.g.*, Luo, 2006)

$$\beta(t) = \sum_{i=1}^{\infty} z_i \int_{t_0}^{t} \phi_i(\tau)\, \mathrm{d}\tau, \tag{8.47}$$

where $z_i \sim \mathrm{N}(0, 1)$, $i = 1, 2, \ldots$ are independent random variables and $\{\phi_i(t)\}$ is an orthonormal basis of the Hilbert space with inner product

$$\langle f, g \rangle = \int_{t_0}^{t_1} f(\tau)\, g(\tau)\, \mathrm{d}\tau. \tag{8.48}$$

The Gaussian random variables are then just the projections of the Brownian motion onto the basis functions:

$$z_i = \int_{t_0}^{t} \phi_i(\tau)\, \mathrm{d}\beta(\tau). \tag{8.49}$$

The series expansion can be interpreted as the following representation for the differential of Brownian motion:

$$\mathrm{d}\beta(t) = \sum_{i=1}^{\infty} z_i\, \phi_i(t)\, \mathrm{d}t. \tag{8.50}$$

We can now consider approximating the following equation by substituting a finite number of terms from the above sum for the term $\mathrm{d}\beta(t)$ into the scalar SDE

$$\mathrm{d}x = f(x,t)\,\mathrm{d}t + L(x,t)\,\mathrm{d}\beta. \tag{8.51}$$

In the limit $N \to \infty$ we could then expect to get the exact solution. However, it has been shown by Wong and Zakai (1965) that this approximation actually converges to the *Stratonovich SDE*

$$\mathrm{d}x = f(x,t)\,\mathrm{d}t + L(x,t) \circ \mathrm{d}\beta. \tag{8.52}$$

That is, we can approximate the above Stratonovich SDE with an equation of the form

$$\mathrm{d}x = f(x,t)\,\mathrm{d}t + L(x,t) \sum_{i=1}^{N} z_i\, \phi_i(t)\,\mathrm{d}t, \tag{8.53}$$

which actually is just an ordinary differential equation

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(x,t) + L(x,t) \sum_{i=1}^{N} z_i\, \phi_i(t), \tag{8.54}$$

and the solution converges to the exact solution, when $N \to \infty$. The solution of an Itô SDE can be approximated by first converting it into the corresponding Stratonovich equation and then approximating the resulting equation.

Now an obvious extension would be to consider a multivariate version of this approximation. Because any multivariate Brownian motion can be formed as a linear combination of independent standard Brownian motions, it is possible to form analogous multivariate approximations. Unfortunately, in the multivariate case the approximation does not generally converge to the Stratonovich solution. There exists basis functions for which this is true (*e.g.*, Haar wavelets), but the convergence is not generally guaranteed.

Another type of series expansion is the so-called *Wiener chaos expansion* (see, *e.g.*, Cameron and Martin, 1947; Luo, 2006). Assume that we indeed are able to solve the Equation (8.54) with any given countably infinite number of values $\{z_1, z_2, \ldots\}$. Then we can see the solution as a function (or functional) of the form

$$x(t) = U(t; z_1, z_2, \ldots). \tag{8.55}$$

The Wiener chaos expansion is the multivariate Fourier–Hermite series for the right hand side above. That is, it is a polynomial expansion of a generic functional of Brownian motion in terms of Gaussian random variables. Hence the expansion is also called *polynomial chaos*.

8.7 Fourier analysis of LTI SDEs

One way to study linear time-invariant SDEs is in Fourier domain. In that case a useful quantity is the spectral density, which is the squared absolute value of the Fourier transform of the process. For example, if the Fourier transform of a scalar process $x(t)$ is $X(i\,\omega)$, then its spectral density is

$$S_x(\omega) = |X(i\,\omega)|^2 = X(i\,\omega)\,X(-i\,\omega), \tag{8.56}$$

In the case of a vector process $\mathbf{x}(t)$ we have the spectral density matrix

$$\mathbf{S}_\mathbf{x}(\omega) = \mathbf{X}(i\,\omega)\,\mathbf{X}^\mathsf{T}(-i\,\omega). \tag{8.57}$$

Now if $\mathbf{w}(t)$ is a white noise process with spectral density $\mathbf{Q}$, it really means that the squared absolute value of the Fourier transform is $\mathbf{Q}$:

$$\mathbf{S}_\mathbf{w}(\omega) = \mathbf{W}(i\,\omega)\,\mathbf{W}^\mathsf{T}(-i\,\omega) = \mathbf{Q}. \tag{8.58}$$

However, one needs to be extra careful when using this, because the Fourier transform of a white noise process is defined only as a kind of limit of smooth processes. Fortunately, as long as we only work with linear systems this definition indeed works. And it provides a useful tool for determining covariance functions of stochastic differential equations.

The covariance function of a zero mean stationary stochastic process $\mathbf{x}(t)$ can be defined as

$$\mathbf{C}_\mathbf{x}(\tau) = \mathrm{E}[\mathbf{x}(t)\,\mathbf{x}^\mathsf{T}(t+\tau)]. \tag{8.59}$$

This function is independent of t, because we have assumed that the process is *stationary*. This means that formally we think that the process has been started at time $t_0 = -\infty$ and it has reached its stationary stage such that its statistic no longer depend on the absolute time t, but only the difference of time steps τ.

The celebrated Wiener–Khinchin theorem says that the covariance function is the inverse Fourier transform of the spectral density:

$$\mathbf{C}_\mathbf{x}(\tau) = \mathcal{F}^{-1}[\mathbf{S}_\mathbf{x}(\omega)]. \tag{8.60}$$

For the white noise process we get

$$\mathbf{C}_\mathbf{w}(\tau) = \mathcal{F}^{-1}[\mathbf{Q}] = \mathbf{Q}\,\mathcal{F}^{-1}[1] = \mathbf{Q}\,\delta(\tau) \tag{8.61}$$

as expected.

Let's now consider the stochastic differential equation

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \mathbf{F}\,\mathbf{x}(t) + \mathbf{L}\,\mathbf{w}(t), \tag{8.62}$$

and assume that it has already reached its stationary stage and hence it also has zero mean. Note that the stationary stage can only exist of the matrix $\mathbf{F}$ corresponds to

a *stable* system, which means that all its eigenvalues have negative real parts. Let's now assume that it is indeed the case.

Similarly as in Section 1.4 we get the following solution for the Fourier transform $\mathbf{X}(i\,\omega)$ of $\mathbf{x}(t)$:

$$\mathbf{X}(i\,\omega) = ((i\,\omega)\,\mathbf{I} - \mathbf{F})^{-1}\;\mathbf{L}\;\mathbf{W}(i\,\omega), \tag{8.63}$$

where $\mathbf{W}(i\,\omega)$ is the *formal* Fourier transform of white noise $\mathbf{w}(t)$. Note that this transform does not strictly exist, because a white noise process is not square-integrable, but let's now pretend that it does. This problem was covered in more detail in Chapter 4.

The spectral density of $\mathbf{x}(t)$ is now given by the matrix

$$\begin{aligned}\mathbf{S}_{\mathbf{x}}(\omega) &= (\mathbf{F} - (i\,\omega)\,\mathbf{I})^{-1}\;\mathbf{L}\;\mathbf{W}(i\,\omega)\;\mathbf{W}^{\mathsf{T}}(-i\,\omega)\,\mathbf{L}^{\mathsf{T}}\;(\mathbf{F} + (i\,\omega)\,\mathbf{I})^{-\mathsf{T}} \\ &= (\mathbf{F} - (i\,\omega)\,\mathbf{I})^{-1}\;\mathbf{L}\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}\;(\mathbf{F} + (i\,\omega)\,\mathbf{I})^{-\mathsf{T}}\end{aligned} \tag{8.64}$$

Thus the covariance function is

$$\mathbf{C}_{\mathbf{x}}(\tau) = \mathcal{F}^{-1}[(\mathbf{F} - (i\,\omega)\,\mathbf{I})^{-1}\;\mathbf{L}\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}\;(\mathbf{F} + (i\,\omega)\,\mathbf{I})^{-\mathsf{T}}]. \tag{8.65}$$

Although this looks complicated, it provides useful means to compute the covariance function of a solution to stochastic differential equation without first explicitly solving the equation.

Note that because $\mathbf{C}_{\mathbf{x}}(0) = \mathbf{P}_\infty$ by Equation (8.93), where $\mathbf{P}_\infty$ is the stationary solution considered in the previous section, we also get the following interesting identity:

$$\mathbf{P}_\infty = \frac{1}{2\pi}\int_{-\infty}^{\infty}(\mathbf{F} - (i\,\omega)\,\mathbf{I})^{-1}\;\mathbf{L}\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}\;(\mathbf{F} + (i\,\omega)\,\mathbf{I})^{-\mathsf{T}}]\,\mathrm{d}\omega, \tag{8.66}$$

which can sometimes be used for computing solutions to stationary (algebraic) Lyapunov equations.

Example 8.1 (Spectrum and covariance of Ornstein–Uhlenbeck)**.** *Let's consider the following scalar SDE (Ornstein–Uhlenbeck process):*

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = -\lambda\,x(t) + w(t), \tag{8.67}$$

where $\lambda > 0$. Taking formal Fourier transform from both sides yields

$$(i\,\omega)\,X(i\,\omega) = -\lambda\,X(i\,\omega) + W(i\,\omega), \tag{8.68}$$

and solving for $X(i\,\omega)$ gives

$$X(i\,\omega) = \frac{W(i\,\omega)}{(i\,\omega) + \lambda}. \tag{8.69}$$

Thus we get the following spectral density

$$S_x(\omega) = \frac{|W(i\,\omega)|^2}{|(i\,\omega) + \lambda|^2} = \frac{q}{\omega^2 + \lambda^2}, \tag{8.70}$$

where q is the spectral density of the white noise input process $w(t)$. The Fourier transform then leads to the covariance function

$$C(\tau) = \frac{q}{2\lambda} \exp(-\lambda\,|\tau|). \tag{8.71}$$

Furthermore we get

$$P_\infty = \frac{1}{2\pi} \int_\infty^\infty \frac{q}{\lambda^2 + \omega^2}\, \mathrm{d}\omega = \frac{q}{2\,\lambda}, \tag{8.72}$$

which indeed is the solution to the stationary Lyapynov equation

$$\frac{\mathrm{d}P}{\mathrm{d}t} = -2\,\lambda\,P + q = 0. \tag{8.73}$$

As pointed out above, the analysis in this section has not been entirely rigorous, because we had to resort to computation of the Fourier transform of white noise

$$W(i\,\omega) = \int_{-\infty}^{\infty} w(t)\, \exp(-i\,\omega\,t)\, \mathrm{d}t, \tag{8.74}$$

which is not well-defined as an ordinary integral. The obvious substitution $\mathrm{d}\beta = w(t)\,\mathrm{d}t$ will not help us either, because we would still have trouble in defining what is meant by this resulting highly oscillatory stochastic process.

The problem can be solved by using the integrated Fourier transform as follows. It can be shown (see, *e.g.*, Van Trees, 1968) that every stationary Gaussian process $x(t)$ has a representation of the form

$$x(t) = \int_0^\infty \exp(i\,\omega\,t)\, \mathrm{d}\zeta(i\,\omega), \tag{8.75}$$

where $\omega \mapsto \zeta(i\,\omega)$ is some complex valued Gaussian process with independent increments. Then the mean squared difference $\mathrm{E}[|\zeta(\omega_{k+1}) - \zeta(\omega_k)|^2]$ roughly corresponds to the mean power on the interval $[\omega_k, \omega_{k+1}]$. The spectral density then corresponds to a function $S(\omega)$ such that

$$\mathrm{E}[|\zeta(\omega_{k+1}) - \zeta(\omega_k)|^2] = \frac{1}{\pi} \int_{\omega_k}^{\omega_{k+1}} S(\omega)\, \mathrm{d}\omega, \tag{8.76}$$

where the constant factor results from two-sidedness of $S(\omega)$ and from the constant factor $(2\pi)^{-1}$ in the inverse Fourier transform.

By replacing the Fourier transform in the above analysis with the integrated Fourier transform, it is possible derive the spectral densities of covariance functions of LTI SDEs without resorting to the formal Fourier transform of white noise.

However, the results remain exactly the same. For more information on this procedure, see, for example, Van Trees (1968).

Another way to treat the problem is to recall that the solution of a LTI ODE of the form

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{F}\,\mathbf{x}(t) + \mathbf{L}\,\mathbf{u}(t), \tag{8.77}$$

where $\mathbf{u}(t)$ is a smooth process, approaches the solution of the corresponding LTI SDE in Stratonovich sense when the correlation length of $\mathbf{u}(t)$ goes to zero. Thus we can start by replacing the formal white noise process with a Gaussian process with covariance function

$$\mathbf{C}_u(\tau;\Delta) = \mathbf{Q}\,\frac{1}{\sqrt{2\pi\,\Delta^2}}\,\exp\left(-\frac{1}{2\,\Delta^2}\tau^2\right) \tag{8.78}$$

which in the limit $\Delta \to 0$ gives the white noise:

$$\lim_{\Delta\to 0}\mathbf{C}_u(\tau;\Delta) = \mathbf{Q}\,\delta(\tau). \tag{8.79}$$

If we now carry out the derivation in Section 8.7, we end up into the following spectral density:

$$\mathbf{S}_\mathbf{x}(\omega;\Delta) = (\mathbf{F} - (i\,\omega)\,\mathbf{I})^{-1}\;\mathbf{L}\,\mathbf{Q}\;\exp\left(-\frac{\Delta^2}{2}\,\omega^2\right)\;\mathbf{L}^\mathsf{T}\;(\mathbf{F} + (i\,\omega)\,\mathbf{I})^{-\mathsf{T}}. \tag{8.80}$$

We can now compute the limit $\Delta \to 0$ to get the spectral density corresponding to the white noise input:

$$\mathbf{S}_\mathbf{x}(\omega) = \lim_{\Delta\to 0}\mathbf{S}_\mathbf{x}(\omega;\Delta) = (\mathbf{F} - (i\,\omega)\,\mathbf{I})^{-1}\;\mathbf{L}\,\mathbf{Q}\,\mathbf{L}^\mathsf{T}\;(\mathbf{F} + (i\,\omega)\,\mathbf{I})^{-\mathsf{T}}, \tag{8.81}$$

which agrees with the result obtained in Section 8.7. This also implies that the covariance function of $\mathbf{x}$ is indeed

$$\mathbf{C}_\mathbf{x}(\tau) = \mathcal{F}^{-1}[(\mathbf{F} - (i\,\omega)\,\mathbf{I})^{-1}\;\mathbf{L}\,\mathbf{Q}\,\mathbf{L}^\mathsf{T}\;(\mathbf{F} + (i\,\omega)\,\mathbf{I})^{-\mathsf{T}}]. \tag{8.82}$$

8.8 Steady state solutions of linear SDEs

In Section (8.7) we considered steady-state solutions of LTI SDEs of the form

$$\mathrm{d}\mathbf{x} = \mathbf{F}\,\mathbf{x}\;\mathrm{d}t + \mathbf{L}\;\mathrm{d}\boldsymbol{\beta}, \tag{8.83}$$

via Fourier domain methods. However, another way of approaching steady state solutions is to notice that at the steady state, the time derivatives of mean and covariance should be zero:

$$\frac{\mathrm{d}\mathbf{m}(t)}{\mathrm{d}t} = \mathbf{F}\,\mathbf{m}(t) = \mathbf{0}, \tag{8.84}$$

$$\frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t} = \mathbf{F}\,\mathbf{P}(t) + \mathbf{P}(t)\,\mathbf{F}^\mathsf{T} + \mathbf{L}\,\mathbf{Q}\,\mathbf{L}^\mathsf{T} = \mathbf{0}. \tag{8.85}$$

The first equation implies that the stationary mean should be identically zero $\mathbf{m}_\infty = \mathbf{0}$. Here we use the subscript ∞ to mean the steady state value which in a sense corresponds to the value after an infinite duration of time. The second equation above leads to so called the Lyapunov equation, which is a special case of so called algebraic Riccati equations (AREs):

$$\mathbf{F}\,\mathbf{P}_\infty + \mathbf{P}_\infty\,\mathbf{F}^\mathsf{T} + \mathbf{L}\,\mathbf{Q}\,\mathbf{L}^\mathsf{T} = \mathbf{0}. \tag{8.86}$$

The steady-state covariance $\mathbf{P}_\infty$ can be algebraically solved from the above equation. Note that although the equation is linear in $\mathbf{P}_\infty$ it cannot be solved via simple matrix inversion, because the matrix $\mathbf{F}$ appears on the left and right hand sides of the covariance. Furthermore $\mathbf{F}$ is not usually invertible. However, most commercial mathematics software (*e.g.*, Matlab) have built-in routines for solving this type of equations numerically.

Let's now use this result to derive the equation for the steady state covariance function of LTI SDE. The general solution of LTI SDE is

$$\mathbf{x}(t) = \exp\left(\mathbf{F}\,(t-t_0)\right)\,\mathbf{x}(t_0) + \int_{t_0}^{t} \exp\left(\mathbf{F}\,(t-\tau)\right)\,\mathbf{L}\,\mathrm{d}\boldsymbol{\beta}(\tau). \tag{8.87}$$

If we let $t_0 \to -\infty$ then this becomes (note that $\mathbf{x}(t)$ becomes zero mean):

$$\mathbf{x}(t) = \int_{-\infty}^{t} \exp\left(\mathbf{F}\,(t-\tau)\right)\,\mathbf{L}\,\mathrm{d}\boldsymbol{\beta}(\tau). \tag{8.88}$$

The covariance function is now given as

$$\begin{aligned}
&\mathrm{E}[\mathbf{x}(t)\,\mathbf{x}^\mathsf{T}(t')] \\
&= \mathrm{E}\left\{\left[\int_{-\infty}^{t} \exp\left(\mathbf{F}\,(t-\tau)\right)\,\mathbf{L}\,\mathrm{d}\boldsymbol{\beta}(\tau)\right]\left[\int_{-\infty}^{t'} \exp\left(\mathbf{F}\,(t'-\tau')\right)\,\mathbf{L}\,\mathrm{d}\boldsymbol{\beta}(\tau')\right]^\mathsf{T}\right\} \\
&= \int_{-\infty}^{\min(t',t)} \exp\left(\mathbf{F}\,(t-\tau)\right)\,\mathbf{L}\,\mathbf{Q}\,\mathbf{L}^\mathsf{T}\,\exp\left(\mathbf{F}\,(t'-\tau)\right)^\mathsf{T}\,\mathrm{d}\tau.
\end{aligned} \tag{8.89}$$

But we already know the following:

$$\mathbf{P}_\infty = \int_{-\infty}^{t} \exp\left(\mathbf{F}\,(t-\tau)\right)\,\mathbf{L}\,\mathbf{Q}\,\mathbf{L}^\mathsf{T}\,\exp\left(\mathbf{F}\,(t-\tau)\right)^\mathsf{T}\,\mathrm{d}\tau, \tag{8.90}$$

which, by definition, should be independent of t. We now get:

- If $t \leq t'$, we have

$$\begin{aligned}
&\mathrm{E}[\mathbf{x}(t)\,\mathbf{x}^{\mathsf{T}}(t')] \\
&= \int_{-\infty}^{t} \exp\left(\mathbf{F}\,(t-\tau)\right)\,\mathbf{L}\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}\,\exp\left(\mathbf{F}\,(t'-\tau)\right)^{\mathsf{T}}\,\mathrm{d}\tau \\
&= \int_{-\infty}^{t} \exp\left(\mathbf{F}\,(t-\tau)\right)\,\mathbf{L}\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}\,\exp\left(\mathbf{F}\,(t'-t+t-\tau)\right)^{\mathsf{T}}\,\mathrm{d}\tau \\
&= \int_{-\infty}^{t} \exp\left(\mathbf{F}\,(t-\tau)\right)\,\mathbf{L}\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}\,\exp\left(\mathbf{F}\,(t-\tau)\right)^{\mathsf{T}}\,\mathrm{d}\tau\,\exp\left(\mathbf{F}\,(t'-t)\right)^{\mathsf{T}} \\
&= \mathbf{P}_\infty\,\exp\left(\mathbf{F}\,(t'-t)\right)^{\mathsf{T}}. \qquad (8.91)
\end{aligned}$$

- If $t > t'$, we get similarly

$$\begin{aligned}
&\mathrm{E}[\mathbf{x}(t)\,\mathbf{x}^{\mathsf{T}}(t')] \\
&= \int_{-\infty}^{t'} \exp\left(\mathbf{F}\,(t-\tau)\right)\,\mathbf{L}\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}\,\exp\left(\mathbf{F}\,(t'-\tau)\right)^{\mathsf{T}}\,\mathrm{d}\tau \\
&= \exp\left(\mathbf{F}\,(t-t')\right) \int_{-\infty}^{t'} \exp\left(\mathbf{F}\,(t-\tau)\right)\,\mathbf{L}\,\mathbf{Q}\,\mathbf{L}^{\mathsf{T}}\,\exp\left(\mathbf{F}\,(t'-\tau)\right)^{\mathsf{T}}\,\mathrm{d}\tau \\
&= \exp\left(\mathbf{F}\,(t-t')\right)\,\mathbf{P}_\infty. \qquad (8.92)
\end{aligned}$$

Thus the stationary covariance function $\mathbf{C}(\tau) = \mathrm{E}[\mathbf{x}(t)\,\mathbf{x}^{\mathsf{T}}(t+\tau)]$ can be expressed as

$$\mathbf{C}(\tau) = \begin{cases} \mathbf{P}_\infty\,\exp\left(\mathbf{F}\,\tau\right)^{\mathsf{T}}, & \text{if } \tau \geq 0 \\ \exp\left(-\mathbf{F}\,\tau\right)\,\mathbf{P}_\infty, & \text{if } \tau < 0. \end{cases} \qquad (8.93)$$

It is also possible to find an analogous representation for the covariance functions of time-varying linear SDEs (Van Trees, 1971).

References

Applebaum, D. (2004). *Lévy Processes and Stochastic Calculus*. Cambridge University Press, Cambridge.

Arasaratnam, I., Haykin, S., and Hurd, T. R. (2010). Cubature Kalman filtering for continuous-discrete systems: Theory and simulations. *IEEE Transactions on Signal Processing*, 58(10):4977–4993.

Archambeau, C. and Opper, M. (2011). Approximate inference for continuous-time Markov processes. In *Bayesian Time Series Models*, chapter 6, pages 125–140. Cambridge University Press.

Åström, K. J. and Wittenmark, B. (1996). *Computer-Controlled Systems: Theory and Design*. Prentice Hall, 3rd edition.

Bar-Shalom, Y., Li, X.-R., and Kirubarajan, T. (2001). *Estimation with Applications to Tracking and Navigation*. Wiley, New York.

Beskos, A., Papaspiliopoulos, O., Roberts, G., and Fearnhead, P. (2006). Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *Journal of the Royal Statistical Society, Series B*, 68(3):333 – 382.

Bucy, R. S. (1965). Nonlinear filtering theory. *IEEE Transactions on Automatic Control*, 10:198–198.

Burrage, K., Burrage, P., Higham, D. J., Kloeden, P. E., and Platen, E. (2006). Comment on "Numerical methods for stochastic differential equations". *Phys. Rev. E*, 74:068701.

Cameron, R. H. and Martin, W. T. (1947). The orthogonal development of non-linear functionals in series of Fourier–Hermite funtionals. *Annals of Mathematics*, 48(2):385–392.

Chaichian, M. and Demichev, A. (2001a). *Path Integrals in Physics Volume 1: Stochastic Processes and Quantum Mechanics*. IoP.

Chaichian, M. and Demichev, A. (2001b). *Path Integrals in Physics Volume 2: Quantum Field Theory, Statistical Physics & Other Modern Applications*. IoP.

Crisan, D. and Rozovskii, B., editors (2011). *The Oxford Handbook of Nonlinear Filtering*. Oxford.

Einstein, A. (1905). Über die von molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Annelen der Physik*, 17:549–560.

Gelb, A. (1974). *Applied Optimal Estimation*. The MIT Press, Cambridge, MA.

Gilsing, H. and Shardlow, T. (2007). SDELab: A package for solving stochastic differential equations in MATLAB. *Journal of Computational and Applied Mathematics*, 205(2):1002–1018.

Girsanov, I. V. (1960). On transforming a certain class of stochastic processes by absolutely continuous substitution of measures. *Theory of Probability and its Applications*, 5:285–

301.

Grewal, M. S. and Andrews, A. P. (2001). *Kalman Filtering, Theory and Practice Using MATLAB*. Wiley, New York.

Hairer, E., Nørsett, S. P., and Wanner, G. (2008). *Solving Ordinary Differential Equations I: Nonstiff Problems*, volume 1 of *Springer Series in Computational Mathematics*. Springer Science & Business.

Ito, K. and Xiong, K. (2000). Gaussian filters for nonlinear filtering problems. *IEEE Transactions on Automatic Control*, 45(5):910–927.

Jazwinski, A. H. (1970). *Stochastic Processes and Filtering Theory*. Academic Press, New York.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME, Journal of Basic Engineering*, 82:35–45.

Kalman, R. E. and Bucy, R. S. (1961). New results in linear filtering and prediction theory. *Transactions of the ASME, Journal of Basic Engineering*, 83:95–108.

Karatzas, I. and Shreve, S. E. (1991). *Brownian Motion and Stochastic Calculus*. Springer-Verlag, New York.

Kloeden, P., Platen, E., and Schurz, H. (1994). *Numerical Solution of SDE Through Computer Experiments*. Springer.

Kloeden, P. E. and Platen, E. (1999). *Numerical Solution to Stochastic Differential Equations*. Springer, New York.

Kreyszig, E. (1993). *Advanced Engineering Mathematics*. John Wiley & Sons, Inc.

Kushner, H. J. (1964). On the differential equations satisfied by conditional probability densities of Markov processes, with applications. *J. SIAM Control Ser. A*, 2(1).

Kushner, H. J. (1967). Approximations to optimal nonlinear filters. *IEEE Transactions on Automatic Control*, 12(5):546–556.

Langevin, P. (1908). Sur la théorie du mouvement brownien (engl. on the theory of Brownian motion). *C. R. Acad. Sci. Paris*, 146:530–533.

Luo, W. (2006). *Wiener Chaos Expansion and Numerical Solutions of Stochastic Partial Differential Equations*. PhD thesis, California Institute of Technology.

Maybeck, P. (1979). *Stochastic Models, Estimation and Control, Volume 1*. Academic Press.

Maybeck, P. (1982). *Stochastic Models, Estimation and Control, Volume 2*. Academic Press.

Nualart, D. (2006). *The Malliavin Calculus and Related Topics*. Springer.

Øksendal, B. (2003). *Stochastic Differential Equations: An Introduction with Applications*. Springer, New York, 6th edition.

Papoulis, A. (1984). *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill.

Rößler, A. (2006). Runge–Kutta methods for Itô stochastic differential equations with scalar noise. *BIT Numerical Mathematics*, 46:97–110.

Rößler, A. (2009). Second order Runge–Kutta methods for Itô stochastic differential equations. *SIAM Journal on Numerical Analysis*, 47(3):1713–1738.

Rößler, A. (2010). Runge–Kutta methods for the strong approximation of solutions of stochastic differential equations. *SIAM Journal on Numerical Analysis*, 48(3):922–952.

Särkkä, S. (2006). *Recursive Bayesian Inference on Stochastic Differential Equations*. Doctoral dissertation, Helsinki University of Technology, Department of Electrical and Communications Engineering.

Särkkä, S. (2007). On unscented Kalman filtering for state estimation of continuous-time nonlinear systems. *IEEE Transactions on Automatic Control*, 52(9):1631–1641.

Särkkä, S. (2013). *Bayesian filtering and smoothing*. Cambridge University Press.

Särkkä, S., Hartikainen, J., Mbalawata, I. S., and Haario, H. (2014). Posterior inference on parameters of stochastic differential equations via non-linear Gaussian filtering and adaptive MCMC. *Statistics and Computing (to appear)*.

Särkkä, S. and Sarmavuori, J. (2013). Gaussian filtering and smoothing for continuous-discrete dynamic systems. *Signal Processing*, 93:500–510.

Särkkä, S. and Solin, A. (2012). On continuous-discrete cubature Kalman filtering. In *Proceedings of SYSID 2012*.

Särkkä, S. and Sottinen, T. (2008). Application of Girsanov theorem to particle filtering of discretely observed continuous-time non-linear systems. *Bayesian Analysis*, 3(3):555–584.

Socha, L. (2008). *Linearization methods for stochastic dynamic systems*. Springer.

Stengel, R. F. (1994). *Optimal Control and Estimation*. Dover Publications, New York.

Stratonovich, R. L. (1968). *Conditional Markov Processes and Their Application to the Theory of Optimal Control*. American Elsevier Publishing Company, Inc.

Tenenbaum, M. and Pollard, H. (1985). *Ordinary Differential Equations*. Dover, New York.

Van Trees, H. L. (1968). *Detection, Estimation, and Modulation Theory Part I*. John Wiley & Sons, New York.

Van Trees, H. L. (1971). *Detection, Estimation, and Modulation Theory Part II*. John Wiley & Sons, New York.

Viterbi, A. J. (1966). *Principles of Coherent Communication*. McGraw-Hill.

Wiktorsson, M. (2001). Joint characteristic function and simultaneous simulation of iterated Itô integrals for multiple independent Brownian motions. *Annals of Applied Probability*, pages 470–487.

Wilkie, J. (2004). Numerical methods for stochastic differential equations. *Physical Review E*, 70, 017701.

Wong, E. and Zakai, M. (1965). On the convergence of ordinary integrals to stochastic integrals. *The Annals of Mathematical Statistics*, 36(5):1560–1564.

Wu, Y., Hu, D., Wu, M., and Hu, X. (2006). A numerical-integration perspective on Gaussian filters. *IEEE Transactions on Signal Processing*, 54(8):2910–2921.

Zakai, M. (1969). On the optimal filtering of diffusion processes. *Zeit. Wahrsch.*, 11:230–243.