ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

IMRT Measurement and Control Laboratory

H. P. Geering, G. Dondi, F. Herzog, S. Keel

Stochastic Systems

April 14, 2011

Contents

1

Probability

In real life, nothing is impossible.
Therefore, say "this event has probability zero"
if you think it is impossible.
Hans P. Geering

A random variable is neither random nor variable.
Gian-Carlo Rota

Probability theory develops the mathematical tools for describing the nature of uncertainty. It is important to note that these tools are deterministic. Randomness only enters when a concrete experiment is made (e.g., we conduct an observation). Since we want to model random phenomena described by random processes and their stochastic differential equations, we need a more rigorous framework than elementary probability theory. This also includes some measure theory. It is not the purpose of this text to rigorously develop measure theory but to provide the reader with the important results (without proofs) and their practical implications. For a more rigorous treatment the reader may refer to [5] or [36].

1.1 Foundations

Definition 1.1. Probability space
A probability space W is a unique triple $W = \{\Omega, \mathcal{F}, P\}$, where Ω is its sample space[1]*, $\mathcal{F}$ its σ-algebra of events*[2]*, and P its probability measure*[3]*.*

The purpose of this section is to clarify the salient details of this very compact definition.

[1] The sample space Ω is the set of all possible samples or elementary events ω: $\Omega = \{\omega \mid \omega \in \Omega\}$.
[2] The σ-algebra $\mathcal{F}$ is the set of all of the considered events A, i.e., subsets of Ω: $\mathcal{F} = \{A \mid A \subseteq \Omega, A \in \mathcal{F}\}$. — See Definition 1.4.
[3] The probability measure P assigns a probability $P(A)$ to every event $A \in \mathcal{F}$: $P : \mathcal{F} \to [0, 1]$. — See Definition 1.10.

The sample space Ω is sometimes called the *universe* of all samples or possible outcomes ω.

Example 1.2. Sample space

- Toss of a coin (with head and tail): $\Omega = \{\mathrm{H}, T\}$.
- Two tosses of a coin: $\Omega = \{HH, HT, TH, TT\}$.
- A cubic die: $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$.
- The positive integers: $\Omega = \{1, 2, 3, \dots\}$.
- The reals: $\Omega = \{\omega \mid \omega \in \mathbb{R}\}$.

Note that the ωs are a mathematical construct and have per se no real or scientific meaning. The ωs in the die example refer to the numbers of dots observed when the die is thrown.

An event A is a subset of Ω. If the outcome ω of the experiment is in the subset A, then the event A is said to have occurred. The set of all subsets of the sample space are denoted by 2^{Ω}. Therefore, the number of all possible events of a finite set is $2^{|\Omega|}$, where $|\Omega| < \infty$ is the number of elements in Ω.

Example 1.3. Events

- Head in the coin toss: $A = \{H\}$.
- Odd number in the roll of a die: $A = \{\omega_1, \omega_3, \omega_5\}$.
- An integer smaller than 5: $A = \{1, 2, 3, 4\}$, where $\Omega = \{1, 2, 3, \dots\}$.
- A real number between 0 and 1: $A = [0, 1]$, where $\Omega = \{\omega \mid \omega \in \mathbb{R}\}$.

We denote the complementary event of A by $A^c = \Omega \backslash A$. When it is possible to determine whether an event A has occurred or not, we must also be able to determine whether A^c has occurred or not. Furthermore, if A and B are events we can also detect the events $A \cap B, A \cup B, A^c \cap B$, etc.

Definition 1.4. σ-algebra
A collection $\mathcal{F}$ of subsets of Ω is called a σ-algebra on Ω if the following properties apply

- *$\Omega \in \mathcal{F}$ and $\emptyset \in \mathcal{F}$ ($\emptyset$ denotes the empty set)*
- *If $A \in \mathcal{F}$ then $\Omega \backslash A = A^c \in \mathcal{F}$:*
  The complementary subset of A is also in Ω.
- *For all $A_i \in \mathcal{F}$: $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$*

The pair $\{\Omega, \mathcal{F}\}$ is called *measure space* and the elements of $\mathcal{F}$ are called *measurable sets*. In our probabilistic environment, a σ-algebra represents all of the events of our experiment. When we define the σ-algebra of an experiment we actually define which events we are able to detect. Therefore, we call them measurable sets or measurable events. A simple example is the roll of a die with incomplete information. If we are only told whether an odd or an even number has been rolled by the die, our σ-algebra would be $\mathcal{F} = \{\emptyset, \{\omega_1, \omega_3, \omega_5\}, \{\omega_2, \omega_4, \omega_6\}, \Omega\}$.

This is the meaning of the σ-algebra when modeling the experiment. Now, after we made a concrete observation, it depends on the σ-algebra how much information we obtained by the observation. Therefore, the σ-algebra is the mathematical construct for modeling informational aspects in an experiment. As a consequence, it determines how much information we get once we conduct some observations.

The simplest or "trivial" σ-algebra is $\{\emptyset, \Omega\}$. With this σ-algebra at hand, we only know that an event has occurred but we have no information about which element ω of Ω has been chosen. In the die example we would only be told that the die has been rolled but not how many eyes showed up.

Conversely, we have full information if $\mathcal{F} = 2^{\Omega}$, i.e., $\mathcal{F}$ consists of all subsets of Ω. This means, we can measure every possible event and therefore we know for every observation which ω in Ω had been chosen. With this σ-algebra as information structure available, there is no more randomness since we have full information. This means for the die example that we know how many eyes showed up once we make an observation.

It is important to note, for the definition of events, that not every subset of Ω is an event. Any subset of Ω, which is not an element of the σ-algebra $\mathcal{F}$, is, mathematically speaking, not an event, and hence does not have a probability. Only the elements of $\mathcal{F}$ are events and have their assigned probabilities.

Example 1.5. σ-algebra of two coin tosses

- $\Omega = \{HH, HT, TH, TT\} = \{\omega_1, \omega_2, \omega_3, \omega_4\}$
- $\mathcal{F}_{min} = \{\emptyset, \Omega\} = \{\emptyset, \{\omega_1, \omega_2, \omega_3, \omega_4\}\}$.
- $\mathcal{F}_{max} = \{\emptyset, \{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_4\}, \{\omega_1, \omega_2\}, \{\omega_1, \omega_3\}, \{\omega_1, \omega_4\}, \{\omega_2, \omega_3\}, \{\omega_2, \omega_4\}, \{\omega_3, \omega_4\}, \{\omega_1, \omega_2, \omega_3\}, \{\omega_1, \omega_2, \omega_4\}, \{\omega_1, \omega_3, \omega_4\}, \{\omega_2, \omega_3, \omega_4\}, \Omega\}$.

The concept of generated σ-algebras is important in probability theory. If, for instance, we are only interested in one subset $A \in \Omega$ in our experiment, the corresponding σ-algebra is $\{\emptyset, A, A^c, \Omega\}$. This leads us to the following definition:

Definition 1.6. $\sigma(\mathcal{C})$: σ-algebra generated by a class $\mathcal{C}$ of subsets
Let $\mathcal{C}$ be a class of subsets of Ω. The σ-algebra generated by $\mathcal{C}$, denoted by $\sigma(\mathcal{C})$, is the smallest σ-algebra $\mathcal{F}$ which includes all elements of $\mathcal{C}$, i.e., $\mathcal{C} \in \mathcal{F}$.

This is actually a very convenient tool for the scientific usage of σ-algebras. If we know what kind of events in experiment we can measure, we denote them by $\mathcal{A}$, then we just work with the σ-algebra generated by $\mathcal{A}$ and we have avoided all the measure-theoretic technicalities for constructing σ-algebras. For the even/odd die example we just consider the σ-algebra generated by $\{\omega_1, \omega_3, \omega_5\}$: $\sigma(\{\omega_1, \omega_3, \omega_5\})$.

If we think of measurability in the engineering context, we think of σ-algebras as the measurable events in an experiment. Therefore, we say that every element $A \in \mathcal{F}$ is $\mathcal{F}$-measurable. The most important σ-algebra used

in this context is the Borel σ-algebra $\mathcal{B}$. The real line $\mathbb{R}$ is often considered as sample space. The Borel σ-algebra is the σ-algebra generated by all open subsets of $\mathbb{R}$ and therefore includes all subsets of $\mathbb{R}$ which are of interest in practical applications.

Definition 1.7. Borel σ-algebra $\mathcal{B}(\mathbb{R})$
The Borel σ-algebra $\mathcal{B}(\mathbb{R})$ is the smallest σ-algebra containing all open intervals in $\mathbb{R}$. The sets in $\mathcal{B}(\mathbb{R})$ are called Borel sets. The extension to the multi-dimensional case, $\mathcal{B}(\mathbb{R}^n)$, is straightforward.

It can be shown that $\mathcal{B}(\mathbb{R})$ contains (for all real numbers a and b):

- open half-lines: $(-\infty, a)$ and (a, ∞),
- union of open half-lines: $(-\infty, a) \cup (b, \infty)$,
- closed interval: $[a, b] = \overline{(-\infty, a) \cup (b, \infty)}$,
- closed half-lines: $(-\infty, a] = \bigcup_{n=1}^{\infty} [a - n, a]$ and $[a, \infty) = \bigcup_{n=1}^{\infty} [a, a + n]$,
- half-open and half-closed $(a, b] = (-\infty, b] \cap (a, \infty)$,
- every set containing only one real number: $\{a\} = \bigcap_{n=1}^{\infty} (a - \frac{1}{n}, a + \frac{1}{n})$,
- every set containing finitely many real numbers: $\{a_1, \cdots, a_n\} = \bigcup_{k=1}^{n} a_k$.

With the Borel σ-algebra $\mathcal{B}(\mathbb{R})$ we are now able to measure events such as $[0, 1]$. This could not be done by just considering the "atomic" elements of Ω, denoted by ω.

What is still needed is the definition of the probability measure itself. Intuitively, we want to assign a probability measure to each event in order to know the frequency of its observation. Mathematically speaking, a measure is some kind of function μ from $\mathcal{F}$ to $\mathbb{R}$. A probability measure has some additional properties.

Definition 1.8. Measure
Let $\mathcal{F}$ be a σ-algebra of Ω and therefore $(\Omega, \mathcal{F})$ be a measurable space. The map

$$\mu : \mathcal{F} \to [0, \infty]$$

is called a measure on $(\Omega, \mathcal{F})$ if μ is countably additive. The measure μ is countably additive (or σ-additive) if $\mu(\emptyset) = 0$, and for every sequence of disjoint sets $(F_i : i \in \mathbb{N})$ in $\mathcal{F}$ with $F = \bigcup_{i \in \mathbb{N}} F_i$ we have

$$\mu(F) = \sum_{i \in \mathbb{N}} \mu(F_i).$$

If μ is countably additive, it is also additive, meaning that for every $F, G \in \mathcal{F}$ we have

$$\mu(F \cup G) = \mu(F) + \mu(G) \quad \text{if and only if} \quad F \cap G = \emptyset \ .$$

The triple $(\Omega, \mathcal{F}, \mu)$ is called a *measure space*.

Intuitively, the measure states that if we take two events which cannot occur simultaneously, then the probability that at least one event occurs is

just the sum of the probabilities of the original events. Note that for example length is a measure on the real line $\mathbb{R}$. This measure is known as the Lebesgue measure.

Definition 1.9. Lebesgue measure on $\mathcal{B}(\mathbb{R})$
The Lebesgue measure on $\mathcal{B}(\mathbb{R})$, denoted by λ, is defined as the measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ which assigns the measure of each interval to be its length.

The Lebesgue measure of a set containing only one point must be zero:

$$\lambda(\{a\}) = 0.$$

The Lebesgue measure of a set containing countably many points ($A = \{a_1, a_2, \cdots\}$) must also be zero:

$$\lambda(A) = \sum_{i=1}^{\infty} \lambda(\{a_i\}) = 0.$$

The Lebesgue measure of a set containing uncountably many points can be either zero, positive and finite, or infinite.

At this point, it is worth noting that there indeed exist subsets on the straight line which do not have a determinable length, e.g., the Vitali sets. But these sets are hard to construct and therefore have no practical importance.

The only problem we are still facing is the range of the measure μ. Our goal is to standardize the probability measure. We do this by defining the probability of the certain event, $\mu(\Omega)$, to have value 1.

We will now state Kolomogorov's axioms (1931) for probability which have generally been accepted.

Definition 1.10. Probability measure
A probability measure P on the sample space Ω with σ-algebra $\mathcal{F}$ is a set function

$$P : \mathcal{F} \to [0, 1],$$

satisfying the following conditions

- $P(\Omega) = 1$.
- *If* $A \in \mathcal{F}$ *then* $P(A) \geq 0$.
- *If* $A_1, A_2, A_3, ... \in \mathcal{F}$ *are mutually disjoint, then*

$$P\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} P(A_i).$$

As a consequence of this definition, we get the following facts:

- $P(\emptyset) = 0$.
- $P(A^c) = 1 - P(A)$ where A^c is the complementary set of A: $A + A^c = \Omega$.

The triple $(\Omega, \mathcal{F}, P)$ is called a *probability space*.

Example 1.11. Finite number of coin tosses
In this experiment, a coin is tossed $n < \infty$ times.

- Sample space: Ω consists of finitely many sample points. Each sample point is a sequence of "Head" (H) and "Tail" (T) with n components: $\omega = (w_1, w_2, \cdots, w_n)$. For $n = 3$ the sample space is:

$$\Omega = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}\,.$$

- σ-algebra: all subsets of Ω (maximal algebra or power set) $\mathcal{A} = 2^{\Omega}$. The total number of subsets in $\mathcal{A}$ is $2^8 = 256$.
- Probability measure: suppose the probability of H on each toss is $P(H) = p$ with $0 \leq p \leq 1$. Then the probability of T is $P(T) = q = 1 - p$. For each $\omega = (w_1, w_2, \cdots, w_n)$ in Ω, we define

$$P(\{\omega\}) = p^{|H \,\in\, \omega|} \cdot q^{|T \,\in\, \omega|}.$$

  The probability of the set $A = \{HHH, HHT, HTH, HTT\} \in \mathcal{A}$ is:

$$\begin{aligned} P(A) = P(\{HHH, HHT, HTH, HTT\}) &= \sum_{\omega \in A} P(\{\omega\}) \\ &= p^3 + p^2 q + p^2 q + p q^2 = p \end{aligned}$$

  This is another way of saying that the probability of H on the first toss is p.

1.2 Random Variables

Consider a sample space Ω which includes all possible outcomes of an experiment. A random variable X assigns a real number to every $\omega \in \Omega$. Glibly speaking, a random variable is just a function from Ω to the real numbers. But X has to be measurable with respect to the σ-algebra of Ω! This is made clearer in the following definition:

Definition 1.12. $\mathcal{F}$-measurable function
The function $f : \Omega \to \mathbb{R}$ defined on $(\Omega, \mathcal{F}, P)$ is called $\mathcal{F}$-measurable if

$$f^{-1}(B) = \{\omega \in \Omega : f(\omega) \in B\} \in \mathcal{F} \quad \textit{for all} \quad B \in \mathcal{B}(\mathbb{R}),$$

i.e., the inverse f^{-1} maps all of the Borel sets $B \subset \mathbb{R}$ to $\mathcal{F}$. — Sometimes, it is easier to work with the following equivalent condition:

$$y \in \mathbb{R} \Rightarrow \{\omega \in \Omega : f(\omega) \leq y\} \in \mathcal{F}\ .$$

The definition of measurable functions is, at first glance, not obvious to understand. If we regard the measurable sets $\mathcal{F}$ as events, an $\mathcal{F}$-measurable function is consistent with the information of the experiment. This means that once we know the (random) value $X(\omega)$ we know which of the events in $\mathcal{F}$ have occurred. Let us consider the easy case of $\mathcal{F} = \{\emptyset, \Omega\}$. For this σ-algebra, only the constant functions are measurable. Consider the equivalence condition of the definition of $\mathcal{F}$-measurable functions. If $f(\omega) = c$ we get, for $y \geq c$, always the whole sample space Ω. Conversely, if $y < c$, by the equivalent condition of the definition of $\mathcal{F}$-measurable functions, we always get the empty set $\emptyset$ since $f(\omega) = c$, for all $\omega \in \Omega$.

For the case of the power set $\mathcal{F} = 2^{\Omega}$ all functions are measurable. We do not need to care about the set $\{\omega \in \Omega : f(\omega) \leq y\}$ for arbitrarily chosen y since every possible subset of Ω is in $\mathcal{F}$. Figure 1.1 gives an overview on the concept of random variables as $\mathcal{F}$-measurable functions. Note that in Figure 1.1 the equivalent condition of Definition 1.12 of $\mathcal{F}$-measurable functions is used.

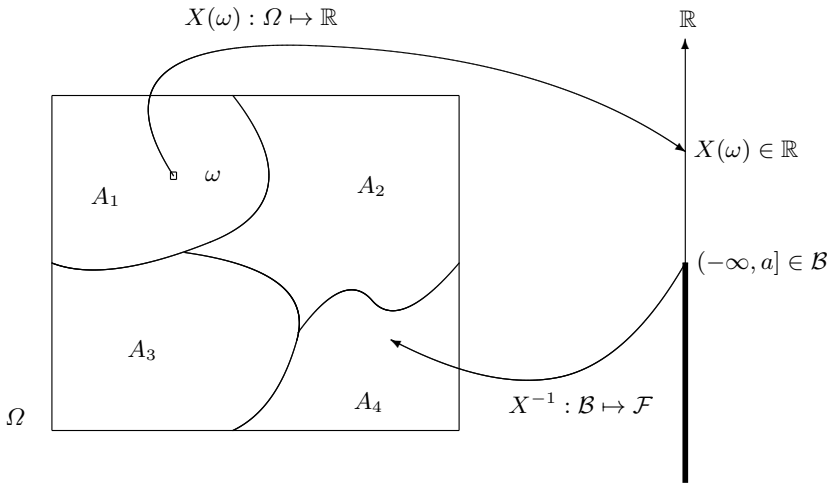

Fig. 1.1. The concept of random variables

For the even/odd die example we consider the following random variable:

$$f(\omega) = \begin{cases} 1 & \text{if } \omega = \omega_1, \omega_2, \omega_3 \\ -1 & \text{if } \omega = \omega_4, \omega_5, \omega_6. \end{cases}$$

This can be thought of as a game where the player wins one Euro when the number of eyes is below 4 and loses one Euro if the number of eyes is above 3. Of course, we already know that this is not a measurable function for $\mathcal{F} = \sigma(\{\omega_1, \omega_3, \omega_5\})$. More formally, we state that the set $\{\omega \in \Omega : f(\omega) \leq -0.51\} = \{\omega_4, \omega_5, \omega_6\} \notin \mathcal{F}$ and therefore f is not $\mathcal{F}$-measurable.

An important example of measurable functions are indicator functions of measurable sets $A \in \mathcal{F}$:

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A. \end{cases}$$

The importance stems from the fact the indicator functions can be used to build more sophisticated functions (such as limits etc.).

Before we state the definition of random variables, we introduce the concept of integration in the stochastic environment. The Lebesgue integral of a function f is a generalization of the Riemann integral, but can be calculated on any sample space Ω. Recall that the integral is just the limit of a sum. Of course, this is also the case for the Lebesgue integral.

Definition 1.13. Lebesgue Integral
Let $(\Omega, \mathcal{F})$ be a measure space, $\mu : \Omega \to \mathbb{R}$ a measure, possibly also taking the values $\pm\infty$, and $f : \Omega \to \mathbb{R}$ an $\mathcal{F}$-measurable function.

- *If f is a simple function, i.e.,*

  $$f(x) = c_i, \quad \text{for all} \quad x \in A_i$$

  where each c_i is a real number and each A_k is a set in $\mathcal{F}$, we define

  $$\int_\Omega f d\mu = \sum_{i=1}^{n} c_i \mu(A_i).$$

- *If f is a nonnegative, measurable but otherwise general function, the construction of the Lebesgue integral is more complicated. The important point here is that we can always construct a sequence of simple functions f_n with $f_n(x) \leq f_{n+1}(x)$ which converges to f:*

  $$\lim_{n\to\infty} f_n(x) = f(x).$$

  With this sequence, the Lebesgue integral is defined by

  $$\int_\Omega f d\mu = \lim_{n\to\infty} \int_\Omega f_n d\mu.$$

- *If f is an arbitrary, measurable function, we have $f = f^+ - f^-$ with*

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

  and then define

  $$\int_\Omega f d\mu = \int_\Omega f^+ dP - \int_\Omega f^- dP.$$

  The integral above may be finite or infinite. It is not defined if $\int_\Omega f^+ dP$ and $\int_\Omega f^- dP$ are both infinite.

As mentioned before, the most important concept of the Lebesgue integral is that it is the limit of approximating sums (as the Riemann-Stieltjes integral is). The Lebesgue integral is more general than the Riemann integral since it is defined over arbitrary sample spaces Ω. Furthermore, the measure μ does not have to be length (as in the Riemann-Stieltjes case). In the important case where $\Omega \equiv \mathbb{R}$, the only difference between the Lebesgue and the Riemann integral is that one is based on the partitioning of the range and the other is based on the partitioning of the domain.

We will take full advantage of the Lebesgue integral when we introduce the concept of expectation.

The Lebesgue integral has all the linearity and comparison properties one would expect. In particular, if $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ are functions and a and b are real constants, then

$$\int_\Omega (aX + bY)\, dP = a \int_\Omega X\, dP + b \int_\Omega Y\, dP\,.$$

If $X(\omega) \leq Y(\omega)$ for all $\omega \in \Omega$, then

$$\int_\Omega X dP \leq \int_\Omega Y\, dP.$$

For quantitative purposes, the definition of the Lebesgue integral is very inconvenient. Finding a convergent sequence of functions is very tedious. But fortunately we have the following theorem:

Theorem 1.14. Riemann-Lebesgue integral equivalence
Let f be a bounded and continuous function on $[x_1, x_2]$ except at a countable number of points in $[x_1, x_2]$. Then both the Riemann and the Lebesgue integral with Lebesgue measure μ exist and are the same:

$$\int_{x_1}^{x_2} f(x)\, dx = \int_{[x_1, x_2]} f\, d\mu.$$

A random variable or random vector is defined as follows:

Definition 1.15. Random variable/vector
A real-valued random variable (vector) X is an $\mathcal{F}$-measurable function defined on a probability space $(\Omega, \mathcal{F}, P)$ mapping its sample space Ω into the real line $\mathbb{R}$ ($\mathbb{R}^n$):

$$X : \Omega \to \mathbb{R}\ (\mathbb{R}^n).$$

Since X is $\mathcal{F}$-measurable we have $X^{-1} : \mathcal{B} \to \mathcal{F}$.

For notational convenience we use $P(X \leq x)$ instead of $P(\{\omega \in \Omega \mid X(\omega) \leq x\})$. As already mentioned, the most important sample space in practice is $\mathbb{R}$ (or $\mathbb{R}^n$). We therefore analyze the case of $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P)$. First, the distribution function is introduced:

Definition 1.16. Distribution function
The distribution function of a random variable X, defined on a probability space $(\Omega, \mathcal{F}, P)$, is defined by:

$$F(x) = P(X(\omega) \leq x) = P(\{\omega \mid X(\omega) \leq x\}).$$

The extension to the multi-dimensional case, $F(x_1, \ldots, x_n)$, is straightforward. The probability measure of the half-open sets in $\mathbb{R}$ is

$$P(a < X \leq b) = P(\{\omega \mid a < X(\omega) \leq b\}) = F(b) - F(a).$$

A close relative of the distribution function is the density function:

Definition 1.17. Density function
The random variable X, defined on a probability space $(\Omega, \mathcal{F}, P)$, has density f with respect to the Lebesgue measure such that f is a non-negative function and for all $A \in \mathcal{F}$:

$$P(\{\omega \mid \omega \in A\}) = \int_A f(x)dx.$$

Again, the extension to the multi-dimensional case, $f(x_1, \ldots, x_n)$, is straightforward.

Example 1.18. Important density functions

- Poisson density or probability mass function ($\lambda > 0$):

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda} \quad , \quad x = 0, 1, 2, \ldots \ .$$

- Multivariate normal density ($x, \mu \in \mathbb{R}^n; \Sigma > 0 \in \mathbb{R}^{n \times n}$):

$$f(x) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \ .$$

- Multivariate t-density with ν degrees of freedom ($x, \mu \in \mathbb{R}^n; \Sigma \in \mathbb{R}^{n \times n}$):

$$f(x) = \frac{\Gamma(\frac{\nu+n}{2})}{\Gamma(\frac{\nu}{2})\sqrt{(\pi\nu)^n \det(\Sigma)}} \Big(1 + \frac{1}{\nu}(x-\mu)^T \Sigma^{-1}(x-\mu)\Big)^{-\frac{1}{2}(\nu+n)} \ .$$

The shorthand notation $X \sim \mathcal{N}(\mu, \sigma^2)$ for normally distributed random variables with parameters μ and σ is often found in the literature. The following properties are useful when dealing with normally distributed random variables:

- If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y = aX + b$, then $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.
- If $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent, then $X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Instead of defining the probability measure on $(\Omega, \mathcal{F})$ in the case of $\Omega \equiv \mathbb{R}$, we can also define the probability measure on the real line $(\mathbb{R}, \mathcal{B})$, which we are far more familiar with from elementary probability theory. Since $\mathcal{F} \in \mathcal{B}(\mathbb{R})$ and X is $\mathcal{F}$-measurable by definition, we can always transform $(\Omega, \mathcal{F})$ to $(\mathbb{R}, \mathcal{B})$ if $\Omega \equiv \mathbb{R}$. This translation is done by the random variable X itself. We therefore limit ourselves to the case of $(\mathbb{R}, \mathcal{B}, \int_{A \in \mathcal{B}} dF)$ because it is a sufficient description of "real world" problems (assuming that the distribution function exists).

Rather than describing a random variable X by its distribution function $F(x)$ or its density function $f(x)$, it is sometimes useful to work with its so-called characteristic function $\varphi(\zeta)$.

Definition 1.19. Characteristic function
For the random variable X with the distribution function F and the density function f, the characteristic function φ is obtained via the following functional transformation:

$$\varphi(\zeta) = \int_{-\infty}^{\infty} e^{j\zeta x}\, dF(x) = \int_{-\infty}^{\infty} e^{j\zeta x} f(x)\, dx \qquad \textit{for all } \zeta \in \mathbb{R}.$$

Notice that the real variable x in the x-domain is replaced by the new real variable ζ in the ζ-domain[4]. — As usual, j denotes $\sqrt{-1}$.

The inverse transformation is:

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-j\zeta x} \varphi(\zeta)\, d\zeta \ .$$

Since we have defined the Lebesgue integral we can now define the expectation and the variance of a random variable in a straight forward manner:

Definition 1.20. Expectation of a random variable
The expectation of a random variable X, defined on a probability space $(\Omega, \mathcal{F}, P)$, is defined by:

$$E[X] = \int_{\Omega} X\, dP = \int_{\mathbb{R}} x\, dF(x) = \int_{\mathbb{R}} x f(x)\, dx.$$

[4] It seems that the poor fellow who invented the characteristic function was not aware of the Fourier transformation, or else he would have chosen $-\zeta$ rather than $+\zeta$ in the exponent of the transformation kernel. — Nevertheless, the nice properties of the Fourier transformation are retained. In particular, convolution of two density functions in the x-domain corresponds to multiplication of their characteristic functions in the ζ-domain.

With this definition at hand, it does not matter what the sample space Ω is. The calculations for the two familiar cases of a finite Ω and $\Omega \equiv \mathbb{R}$ with continuous random variables remain the same.

More generally, the expectation of an arbitrary function g of a random variable X is defined as

$$E[g(X)] = \int_{\Omega} g(X) dP.$$

Definition 1.21. Variance of a random variable
The variance of a random variable X, defined on a probability space $(\Omega, \mathcal{F}, P)$, *is defined by:*

$$\mathrm{var}(X) = \sigma^2(X) = E[(X - E[X])^2] = \int_{\Omega} (X - E[X])^2 \, dP = E[X^2] - E[X]^2.$$

The square root of the variance, σ, is called the standard deviation. The concept of (in-)dependence of random variables is an important topic in probability. Calculations and reasoning are a lot easier once we know that two random variables are independent.

Definition 1.22. Independence of random variables
The random variables $X_1, X_2, \ldots, X_n$ *are independent if*

$$P\Big(\bigcap_{i=1}^{n} \{X_i \in A_i\}\Big) = \prod_{i=1}^{n} P(\{X_i \in A_i\}) \quad \textit{for all} \quad A_i \in \mathcal{F}.$$

As an important consequence, this yields

$$E\Big[\prod_{i=1}^{n} X_i\Big] = \prod_{i=1}^{n} E[X_i]$$

for independent random variables.

If we assume that $f_i(x_i)$ is the density of the random variable X_i, then the independence condition is equivalent to

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_i(x_i) \ .$$

For two random variables we define their covariance to be

$$\mathrm{cov}(X_1, X_2) = E\big[(X_1 - E[X_1])(X_2 - E[X_2])\big],$$

and the correlation coefficient ρ

$$\rho(X_1, X_2) = \frac{\mathrm{cov}(X_1, X_2)}{\sigma(X_1)\sigma(X_2)} \in [-1, 1] \ .$$

It is important to notice that uncorrelated random variables need not be independent.

1.3 Conditional Expectation

The concept of conditional expectation is very important because it plays a fundamental role in many applications of probability. As we already know from fundamental probability theory, conditional probability makes use of additional information. For the discrete case, the probability of A, given B, is

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} \quad , \quad P(B) > 0.$$

This formula is also known as Bayes' rule. Intuitively, this formula is very simple if we look at it in the following way: since we know for sure that $\omega \in B$, it is natural to consider B as our new sample space $\tilde{\Omega}$. Therefore, we only need to scale $P(A \cap B)$ by $1/P(B)$ in order to have $P(\tilde{\Omega} = B) = 1$.

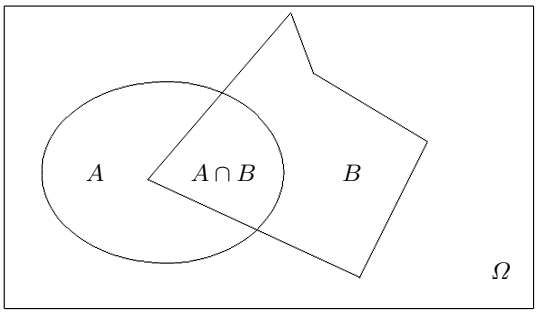

Fig. 1.2. The concept of conditional expectation

From the conditional probability we get conditional expectation of the random variable Y, given B, as

$$E(Y|B) = \frac{E(XI_B)}{P(B)} \quad , \quad P(B) > 0.$$

where I_B denotes the indicator function of the set B.

We have considered the set B above as an event and we have introduced the σ-algebra $\mathcal{F}$ as the collection of "measurable" events. Therefore, the natural extension for the conditional expectation is the inclusion of the σ-algebra, generated by a random variable or vector. The concept for a discrete random variable is rather simple. We consider the sets where the random variable X which takes distinct values x_i. We consider the sets $A_i = \{\omega \mid X(\omega) = x_i\}$ which, together, are a disjoint partition of Ω. We then use the concept of generated σ-algebras. Choose $\mathcal{C} = \{A_1, A_2, \dots\}$ and call $\sigma(\mathcal{C}) = \sigma(X)$ the σ-algebra generated by X. In this setup, we define the conditional expectation of the random variable Y, given the value of the random variable X, to be

$$E(Y|X = x_i) = E(Y|\sigma(X)).$$

Note that the values x_i do not matter for the conditional expectation. Rather, the sets $A_i = \{\omega \mid X(\omega) = x_i\}$ determine the conditional expectation.

Example 1.23. Simple die game
Consider a game where a die is rolled: $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$. The player wins one Pound Sterling when the number of eyes is even and loses one Pound Sterling if the number is odd. Therefore, the random variable Y of the player's win or loss is

$$Y(\omega) = \begin{cases} 1 & \text{if } \omega = \omega_2, \omega_4, \omega_6 \\ -1 & \text{if } \omega = \omega_1, \omega_3, \omega_5. \end{cases}$$

Consider another random variable X on Ω which indicates whether the number is above three or below four:

$$X(\omega) = \begin{cases} 0 & \text{if } \omega = \omega_1, \omega_2, \omega_3 \\ 1 & \text{if } \omega = \omega_4, \omega_5, \omega_6. \end{cases}$$

We want to compute the conditional expectation of Y if we know the value of X. The σ-algebra generated by X is $\sigma(\{\omega_1, \omega_2, \omega_3\})$. This yields for the conditional expectation

$$E(Y|X) = \begin{cases} -\frac{1}{3} & \text{if } \omega \in \{\omega_1, \omega_2, \omega_3\} \text{ or } X(\omega) = 0, \text{ respectively} \\ \frac{1}{3} & \text{if } \omega \in \{\omega_4, \omega_5, \omega_6\} \text{ or } X(\omega) = 1, \text{ respectively.} \end{cases}$$

Note that the actual value of X does not influence the value of the conditional expectation.

We now want to extend the conditional expectation for the general case of a probability space $(\Omega, \mathcal{F}, P)$. As already mentioned, the mathematical construct for describing additional information are σ-algebras. The definition of the conditional expectation is:

Definition 1.24. Conditional expectation
Let X be a random variable defined on the probability space $(\Omega, \mathcal{F}, P)$ with $E[|X|] < \infty$. Furthermore, let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$ ($\mathcal{G} \subseteq \mathcal{F}$). Then there exists a random variable Y with the following properties:

1. *Y is $\mathcal{G}$-measurable.*
2. *$E[|Y|] < \infty$.*
3. *For all sets G in $\mathcal{G}$ we have*

$$\int_G Y \, dP = \int_G X \, dP \qquad \text{for all} \quad G \in \mathcal{G}.$$

The random variable $Y = E[X|\mathcal{G}]$ is called conditional expectation.

It can be shown that if another random variable Z satisfies the conditions above we have $Z = Y$ almost surely.

At first glance, this definition seems very unpleasant but it is not that bad. The most obvious fact is that $Y = E[X|\mathcal{G}]$ is constant on all of the sets in

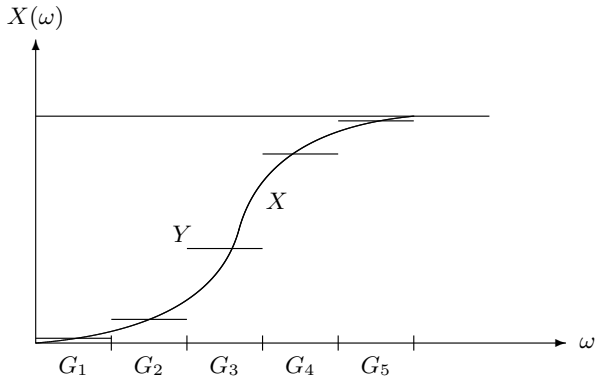

Fig. 1.3. Conditional expectation as piecewise linear approximation

$\mathcal{G}$. Therefore, Y is a piecewise linear approximation of X. This is shown in Figure 1.3. The sample space is partitioned into five mutually disjoint subsets: $\Omega = G_1 \cup G_2 \cup G_3 \cup G_4 \cup G_5$. Then Y is just a coarser version of X.

It is easily seen that the conditional expectation for the trivial σ-algebra $\{\emptyset, \Omega\}$ equals the unconditional expectation $Y = E[X|\{\emptyset, \Omega\}] = \int_\Omega X dP = E[X]$. — Some useful properties of the conditional expectation are stated below:

Property 1.25. Conditional expectation

- $E(E(X|\mathcal{F})) = E(X)$.
- If X is $\mathcal{F}$-measurable, then $E(X|\mathcal{F}) = X$.
- Linearity: $E(\alpha X_1 + \beta X_2|\mathcal{F}) = \alpha E(X_1|\mathcal{F}) + \beta E(X_2|\mathcal{F})$.
- Positivity: If $X \geq 0$ almost surely, then $E(X|\mathcal{F}) \geq 0$.
- Tower property: If $\mathcal{G}$ is a sub-σ-algebra of $\mathcal{F}$, then

$$E(E(X|\mathcal{F})|\mathcal{G}) = E(X|\mathcal{G}).$$

- Taking out what is known: If Z is $\mathcal{G}$-measurable, then

$$E(ZX|\mathcal{G}) = Z \cdot E(X|\mathcal{G}).$$

From elementary probability theory we already know the conditional density. For two random variables X_1 and X_2 which have the joint density function $f(x_1, x_2)$, the marginal density of X_1 is defined by

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_2.$$

The conditional density of X_2, given $X_1 = x_1$ is given by

$$f(x_2 \mid X_1 = x_1) = \frac{f(x_1, x_2)}{f_{X_1}(x_1)}.$$

1.4 Convergence of Random Variables

The best known convergence property in probability is the law of large numbers. Loosely speaking, the law of large numbers states that the probability of an event A can be determined arbitrarily precisely by making sufficiently many observations. This fact was used long before Kolmogorov's axiomatic definitions of probability.

There are four convergence concepts which will be discussed in this section. We consider a sequence of random variables $\{X_n\}$ and a random variable X, all of them defined on the probability space $(\Omega, \mathcal{F}, P)$.

1. The sequence $\{X_n\}$ converges to X with probability one (or almost surely), $X_n \xrightarrow{1} X$, if
$$P(\{\omega \in \Omega \mid \lim_{n\to\infty}(X_n(\omega)) = X(\omega)\}) = 1.$$
This means that X_n converges to X in the usual sense except for null sets of Ω.
2. The sequence $\{X_n\}$ converges to X in probability, $X_n \xrightarrow{p} X$, if
$$\lim_{n\to\infty}\Big(P(\{\omega \in \Omega \mid |X_n(\omega) - X(\omega)| > \varepsilon\})\Big) = 0, \quad \text{for all} \quad \varepsilon > 0.$$
3. The sequence $\{X_n\}$ converges to X in L^p, $X_n \xrightarrow{L^p} X$, if
$$\lim_{n\to\infty}\Big(E(|X_n(\omega) - X(\omega)|^p)\Big) = 0.$$
4. The sequence $\{X_n\}$ converges to X in distribution, $X_n \xrightarrow{d} X$, if
$$\lim_{n\to\infty} F_n(x) = F(x), \quad \text{for all} \quad x \in \mathbb{R},$$
where F_n denotes the distribution function of X_n and F denotes the distribution function of X.

Obviously, the different convergence concepts are not independent of each other. Figure 1.4 summarizes the dependence of the different types of convergence.

The upper right corner of Figure 1.4 states that if a sequence converges in L^p then it also converges in L^q for all $q < p$. The most important case is convergence in the mean-square sense. From the results in this section we therefore only have to check convergence in L^2 in order to have also convergence in L^1.

In general, we cannot compare almost sure convergence and convergence in L^p. Nevertheless, both types of convergence imply convergence in probability. Note that almost sure convergence is usually hard to prove whereas convergence in L^p is usually a lot easier to prove.

The weakest concept of convergence considered here is convergence in distribution. This concept only describes the statistical properties of the limit of the sequence.

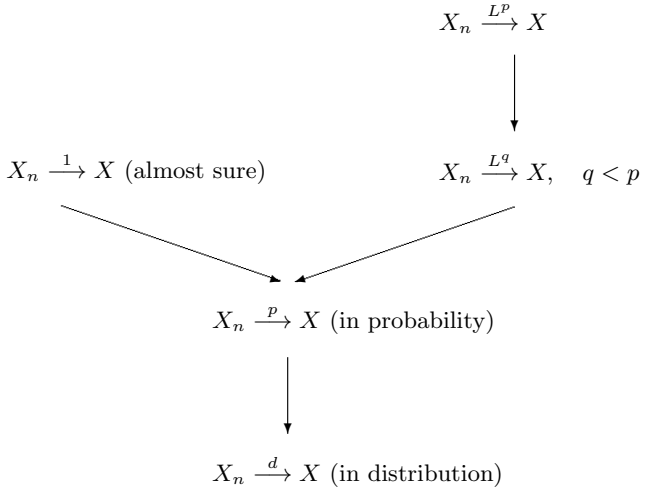

Fig. 1.4. Convergence of random variables

Notes and Comments

Besides the rigorous treatments in [36], [5], [7], or [24], there are very readable textbooks on the subject of this chapter. Among them are [26], [3], and [11].

1.5 Exercises

1. A fair six-faced die is thrown repetitively. What is the probability that in the first ten throws you always get six eyes? What is the probability that you will always get six eyes in the throws eleven through twenty as well?
2. You are attending an entertainment show. On the stage, there are three doors. Behind one of them, there is a goat. If you can guess correctly, behind which one it is, you can keep it. — You make an initial guess about the door but you do not tell anybody. The showmaster does not know your guess, and he does not know where the goat is. He opens one door at random. It's not the door you have chosen, and the goat is not there. — Now you must tell the showmaster which door he should open. In order to maximize the winning probability, do you stick with your initial guess or do you switch to the other door? — Hint: Start with a stochastic simulation ....
3. We have two independent real random variables x_1 and x_2 with the density functions f_1 and f_2, respectively. Show that the density function f of the sum $x_1 + x_2$ is obtained by the convolution of f_1 and f_2.
4. Who invented the *characteristic function* (Definition 1.19)?
5. Verify that the convolution of densities, $f = f_1 * f_2$, corresponds to the multiplication of their characteristic functions: $\varphi(f) = \varphi(f_1) \cdot \varphi(f_2)$.

6. The central limit theorem of probability theory says that the sum (and the average) of independent and identically distributed real random variables converges to a random variable with a Gaussian distribution. What is the implication of this when we are working with characteristic functions? Choose an example and verify!

2

Random Processes

She: What is white noise?
He: It is the best model of a totally unpredictable process.
She: Are you implying, I am white noise?
He: No, it does not exist.
Dialogue of an unknown couple

2.1 Introduction

In the first chapter, we have introduced the mathematical framework to describe random observations. This chapter extends these concepts with an additional time dependence component. In order to model randomness in signals (noise signals), we introduce the notion of random processes.

Once again we want to stress the fact that the tools are deterministic mathematical constructs; randomness only enters when observations are conducted. We first state the classic definition of random processes.

Definition 2.1. Random process
A random (or stochastic) process $\{X_t, t \in T\}$ is a collection of random variables on the same probability space $(\Omega, \mathcal{F}, P)$. The index set T is usually representing time and can be either an interval $[t_1, t_2]$ or a discrete set. Therefore, the random process X can be written as a function:

$$X : \mathbb{R} \times \Omega \to \mathbb{R}, \quad (t, \omega) \mapsto X(t, \omega)$$

In the stochastic interpretation, a sample ω is chosen from the sample space Ω "at random". This yields the "stochastic signal" or "noise signal" $r(\cdot, \omega)$ defined on the index set T. This signal is also denoted as sample path, realization, or trajectory.

Remark 2.2. Notation
We introduced random or stochastic processes as functions with two arguments: t and ω. We will, however, omit the argument ω for brevity as it is done in most text books: $X(t, \omega) = X(t)$.

By the definition of random processes, we know that the amount of information is increasing with time. Again, we need the concept of sigma algebras. We assume that information is not lost with increasing time and therefore the corresponding σ-algebras will increase over time as more and more information becomes available. This concept is called filtration.

Definition 2.3. Filtration/adapted process
A collection $\{\mathcal{F}_t\}_{t\geq 0}$ of sub σ-algebras is called filtration if, for every $s \leq t$, we have $\mathcal{F}_s \subseteq \mathcal{F}_t$. The random variables $\{X_t : 0 \leq t \leq \infty\}$ are called adapted to the filtration $\mathcal{F}_t$ if, for every t, X_t is measurable with respect to $\mathcal{F}_t$.

The concept of filtration is easily understood with a simple example.

Example 2.4. Suppose we have a sample space of four elements: $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$. At time zero, we do not have any information about which ω has been chosen. At time $\frac{T}{2}$ we know whether we have $\{\omega_1, \omega_2\}$ or $\{\omega_3, \omega_4\}$. At time T, we have full information.

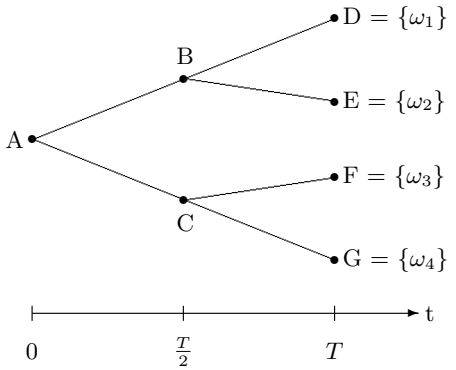

Fig. 2.1. Example of a filtration

Therefore, we have the following σ-algebras:

$$\mathcal{F}_t = \begin{cases} \{\emptyset, \Omega\}, & t \in [0, \frac{T}{2}) \\ \{\emptyset, \{\omega_1, \omega_2\}, \{\omega_3, \omega_4\}, \Omega\}, & t \in [\frac{T}{2}, T) \\ \mathcal{F}_{max} = 2^{\Omega}, & t = T. \end{cases}$$

Thus, $\mathcal{F}_0$ represents initial information whereas $\mathcal{F}_\infty$ represents full information (all we will ever know). Therefore, a stochastic process is said to be defined on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, P)$.

Before going into the topics of random processes, stationary random processes, Gaussian random processes, etc., let us first recapitulate two (almost) trivial properties of deterministic functions:

Let $x(\cdot)$ be a real, continuously differentiable function defined on the interval $[0, T]$. Its continuous differentiability implies both a bounded total variation and a vanishing "sum of squared increments":

1. Total variation:
$$\int_0^T \left| \frac{dx(t)}{dt} \right| dt < \infty$$

2. "Sum of squares":

$$\lim_{N\to\infty}\sum_{k=1}^{N}\left(x\Big(k\frac{T}{N}\Big)-x\Big((k-1)\frac{T}{N}\Big)\right)^2=0$$

Random processes do not have either of these nice smoothness properties in general. This allows the desired "wild" and "random" behavior of the (sample) "noise signals".

2.2 Classes of Processes

2.2.1 Markov Process

A Markov process X is a particular type of stochastic process where only the present value $X(t)$ is relevant for predicting the future evolution of X. Therefore, the past and the future of a Markov process have no direct interconnection. More formally we have:

Definition 2.5. Markov process
A continuous-time stochastic process $X(t),\ t\in T$, is called a Markov process if for any finite parameter set $\{t_i : t_i < t_{i+1}\}\in T$ we have

$$P(X(t_{n+1})\in B \mid X(t_1),\ldots,X(t_n)) = P(X(t_{n+1})\in B \mid X(t_n))\ .$$

For a Markov process $X(t)$ we define the *transition probability*, denoted by $P(s,x,t,B)$, as follows:

$$\mathbf{P}(s,x,t,B) = P(X(t)\in B \mid X(s)=x) \quad , \quad 0\le s<t.$$

The function $\mathbf{P}$ gives the probability of $X(t)$ lying in the set B at time t, given the value x of the process at time s. The transition density $\mathbf{p}$ is implicitly defined as

$$\mathbf{P}(s,x,t,B)=\int_B \mathbf{p}(s,x,t,y)\ dy.$$

2.2.2 Gaussian Process

A stochastic process is called Gaussian if all of its joint probability distributions are Gaussian. If $X(t)$ is a Gaussian process, then $X(t)\sim\mathcal{N}(\mu(t),\sigma^2(t))$ for all t, where $\mu(t)$ and $\sigma^2(t)$ are arbitrary functions. A Gaussian process is fully characterized by its mean and covariance function. Gaussian processes do have many nice mathematical properties. For example performing linear algebraic operations on a Gaussian process yields a Gaussian process. Another important property is that the limit of a Gaussian random sequence remains a Gaussian process. Hence, the mean square derivatives and integrals of Gaussian processes are Gaussian processes themselves. These crucial properties will be needed later on.

2.2.3 Martingales

A stochastic process $X(t)$ is a martingale on the filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, P)$ if the following conditions hold:

- $X(t)$ is $\{\mathcal{F}_t\}_{t\geq 0}$-adapted, $E[|X(t)|] < \infty$ for all $t \geq 0$.
- $\mathrm{E}[X(t)|\mathcal{F}_s] = \mathrm{X(s)}$ a.s. for all $s \in [0, t]$.

From this definition, it follows that the best prediction of a martingale process is its current value. We therefore state that martingale processes model fair games. If we consider a coin tossing game where the player gains one dollar on head and loses one dollar on tail the wealth of the player follows a martingale. The martingale theory is a fundamental tool in finance, and the theory behind it is vast.

2.2.4 Diffusions

A diffusion is a Markov process with continuous trajectories such that for each time t and state $X(t)$ the following limits exist

$$\mu(t, X(t)) := \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} E[X(t + \Delta t) - X(t)|X(t)],$$
$$\sigma^2(t, X(t)) := \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} E[\{X(t + \Delta t) - X(t)\}^2|X(t)].$$

For these limits, $\mu(t, X(t))$ is called drift and $\sigma^2(t, X(t))$ is called the diffusion coefficient. Since diffusions are Markov processes we expect a relationship between the transition probability and $\mu(t, X(t))$, $\sigma^2(t, X(t))$. Actually, under certain assumptions, the transition probability is uniquely determined by $\mu(t, X(t))$ and $\sigma^2(t, X(t))$. This is a pretty surprising result because usually a distribution is not completely determined by its first two moments.

2.3 Brownian Motion and White Noise

2.3.1 Brownian Motion

Motivated by the apparently random walk of a tiny particle in a fluid (observed by the Scottish botanist Robert Brown in 1827), the American mathematician Norbert Wiener stipulated the following assumptions for a stationary random process $W(\cdot, \cdot)$ with independent increments in 1923:

Definition 2.6. Brownian motion
A stochastic process W(t) is called Brownian motion if

1. *Independence: $W(t+\Delta t) - W(t)$ is independent of $\{W(\tau)\}$ for all $\tau \leq t$.*
2. *Stationarity: The distribution of $W(t + \Delta t) - W(t)$ does not depend on t.*

3. Continuity: $\lim_{\Delta t \downarrow 0} \dfrac{P(|W(t+\Delta t) - W(t)| \geq \delta)}{\Delta t} = 0$ *for all* $\delta > 0$.

Please note that the third assumption is expressed with probabilities: discontinuities in sample functions can only occur with probability zero. Hence, there is a version of the Brownian motion with *all* sample functions continuous. (This technicality is not of any practical importance.)

This definition induces the distribution of the process W_t:

Theorem 2.7. Normally distributed increments of Brownian motion
If $W(t)$ *is a Brownian motion, then* $W(t) - W(0)$ *is a normal random variable with mean* μt *and variance* $\sigma^2 t$*, where* μ *and* σ *are constant real numbers.*

As a result of this theorem, we have the following density function of a Brownian motion:

$$f_{W(t)}(x) = \frac{1}{\sqrt{2\pi\sigma^2 t}} e^{-\frac{(x-\mu t)^2}{2\sigma^2 t}}.$$

An irritating property of Brownian motion is that its sample paths are not differentiable. This is easily verified in the mean-square sense:

$$E\Big[\Big(\frac{W(t+\Delta t) - W(t)}{\Delta t}\Big)^2\Big] = \frac{E[(W(t+\Delta t) - W(t))^2]}{\Delta t^2} = \frac{\sigma^2}{\Delta t}.$$

This diverges for $\Delta t \to 0$ and therefore it is not differentiable in L^2. This is also the case for almost sure convergence, but this is much more difficult to prove.

The Brownian motion has many more bizarre and intriguing properties. Some of them are listed below:

- Autocovariance function: $E\{(W(t) - \mu t)(W(\tau) - \mu\tau)\} = \sigma^2 \min(t, \tau)$
- $\mathrm{Var}\left\{\dfrac{W(t)}{t}\right\} = \dfrac{\sigma^2}{t}$
- $\lim\limits_{t\to\infty} \dfrac{W(t) - \mu t}{t} = 0$ with probability 1
- The total variation of the Brownian motion over a finite interval $[0, T]$ is infinite!
- The "sum of squares" of a drift-free Brownian motion is deterministic:
  $\lim\limits_{N\to\infty} \sum\limits_{k=1}^{N} \left(W\Big(k\frac{T}{N}\Big) - W\Big((k-1)\frac{T}{N}\Big)\right)^2 = \sigma^2 T$
- Infinite oscillations:
  Let Y_0, Y_1, ... be mutually independent random variables with identical normal distributions $\mathcal{N}(0, 1)$. The random process
  $X(t) = \dfrac{Y_0}{\sqrt{\pi}}\, t + \sum\limits_{k=1}^{\infty} \sqrt{\dfrac{2}{\pi}} \dfrac{Y_k}{k} \sin kt \quad$ for $t \in [0, \pi]$
  is a normalized Brownian motion on the interval $[0, \pi]$.

- If $W(\cdot)$ is a Brownian motion on the interval $[0, \infty)$, then the following process $W^*(\cdot)$ is a Brownian motion as well:
  $W^*(t) = \begin{cases} tW(\frac{1}{t}), & \text{for } t > 0; \\ 0, & \text{for } t = 0. \end{cases}$
- Zero crossings:
  In a finite interval $[0, T]$, every sample of a drift-free Brownian motion has infinitely many zero-crossings. The set of zero-crossings is dense in $[0, T]$, i.e., no sample path has isolated zero-crossings!

Definition 2.8. Standard Brownian motion
A Brownian motion is standard if

$$\begin{aligned} W(0) &= 0 \quad a.s., \\ E[W(t)] &= 0 \quad (\mu = 0), \\ E[W^2(t)] &= t \quad (\sigma^2 = 1). \end{aligned}$$

Note that Brownian motion is usually assumed to be standard if not explicitly stated otherwise.

We have already stated that the "sum of squares" of a drift-free Brownian motion is deterministic. This can be formulated more generally as follows:

Theorem 2.9. Quadratic variation of standard Brownian motion
The quadratic variation of standard Brownian motion over $[0, t]$ *exists and equals* t*. Formally, we can also write* $(dW(t))^2 = dt$*.*

2.3.2 White Noise

As we have seen in Section 2.3.1, a Brownian motion is continuous but nowhere differentiable.

Nevertheless, in engineering circles, it is customary to define a random process $v(\cdot)$ called *stationary white noise* as the formal derivative of a general Brownian motion $W(\cdot)$ with the drift parameter μ and the variance parameter σ^2:

$$v(t) = \frac{dW(t)}{dt} \quad .$$

Usually, the "initial" time is shifted from $t = 0$ to $t = -\infty$. In this way, the white noise $v(\cdot)$ becomes truly stationary on the infinite time interval $(-\infty, \infty)$. Without loss of generality, we may assume that $v(t)$ is Gaussian for all t.

This stationary white noise is characterized uniquely as follows:

- Expected value:
  $$E\{v(t)\} \equiv \mu$$
- Autocovariance function:
  $$\Sigma(\tau) = E\{[v(t+\tau) - \mu][v(t) - \mu]\} \equiv \sigma^2 \delta(\tau)$$

- Spectral density function:

$$S(\omega) = \mathcal{F}\{\Sigma(\tau)\} = \int_{-\infty}^{\infty} e^{-j\omega\tau}\,\Sigma(\tau)\,d\tau \equiv \sigma^2 \quad .$$

Of course, the characterizations by the autocovariance function *and* the spectral density function are redundant.

Using white noise as the model of a completely unpredictable random process, we can say: the continuous-time measurement y of the third state variable x_3 is corrupted by an additive white noise v:

$$y(t) = x_3(t) + v(t) \quad .$$

Expressing the same fact in full mathematical correctness using a Brownian motion, we would have to say: The integral of the continuous-time measurement y of the third state variable x_3 is corrupted by an additive Brownian motion W:

$$\int_0^t y(t)\,dt = \int_0^t x_3(t)\,dt + W(t) \quad .$$

Yet another way of expressing ourselves in full mathematical correctness could be: The short-time averaged (or smoothed) measurement $\overline{y}$ of the third state variable x_3 is corrupted by an additive increment of a Brownian motion W:

$$\overline{y}(t) = \frac{1}{\Delta T}\int_{t-\Delta T}^{t} y(t)\,dt = \frac{1}{\Delta T}\int_{t-\Delta T}^{t} x_3(t)\,dt + \frac{W(t) - W(t-\Delta T)}{\Delta T} \quad .$$

It should be obvious where this leads to mathematically as $\Delta T \downarrow 0$.

Of course, smoothing by averaging is not optimal. Rather, a Kalman filter (or extended Kalman filter) should be used. (See Chapter 4.)

The Brownian motion W on the time interval $[0, \infty)$ can be retrieved from the stationary white noise v by integration:

$$W(t) = \int_0^t v(\alpha)\,d\alpha \quad .$$

Mathematicians prefer to write this equation in the following way:

$$W(t) = \int_0^t v(\alpha)\,d\alpha = \int_0^t \frac{dW(\alpha)}{d\alpha}\,d\alpha = \int_0^t dW(\alpha) \quad .$$

Consequently, a Brownian motion X with the drift parameter μ, the variance parameter σ^2, and the initial time $t = 0$ satisfies the following stochastic differential equation, where W is a standard Brownian motion:

$$\begin{aligned} dX(t) &= \mu dt + \sigma dW(t) \\ X(0) &= 0 \quad . \end{aligned}$$

2.3.3 Generalizations

Defining the Brownian motion via a stochastic differential equation involving the drift parameter μ and the volatility parameter σ leads to the following rather straightforward generalizations:

- Instationary Brownian motion:

$$dY(t) = \mu(t)dt + \sigma(t)dW(t) \quad .$$

- Locally Brownian motion:

$$dY(t) = \mu(Y(t), t)dt + \sigma(Y(t), t)dW(t) \quad .$$

- Geometric Brownian motion:

$$dY(t) = \mu Y(t)dt + \sigma Y(t)dW(t) \quad .$$

  This is a special case of a locally Brownian motion. Note that both its drift parameter $\mu Y(t)$ and its volatility parameter $\sigma Y(t)$ are proportional to the value $Y(t)$ of the random process. This model is very popular and useful in the area of finance.
- Ornstein-Uhlenbeck process or exponentially correlated noise:

$$dY(t) = -aY(t)dt + b\sigma dW(t) \qquad \text{with } a > 0 \quad .$$

2.4 Poisson Processes

In the previous section, the Wiener process or Brownian motion has been introduced. Brownian motion is a stochastic process in continuous-time with continuous realizations. In this section, we introduce a stochastic process in continuous time with discontinuous realizations. A suitable stochastic model for this kind of behavior is a Poisson process. Often, these discontinuities in financial time series are called "extreme" or "rare" events. For example, the drop of the Dow Jones Index of 22.6% on October 19, 1987 constitutes such a "rare" event. To account for such a large drop in the time series, Brownian motion is not a sufficient model and thus there is a need to describe discontinuous stochastic processes.

Definition 2.10. Poisson process
A Poisson process with parameter λ is a collection of random variables $Q(t),\ t \in [0, \infty)$ defined on $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, P)$ having the discrete state space $N = \{0, 1, 2, \ldots\}$ and satisfying the following properties:

1. *Q(0)=0 with probability one.*
2. *For each $0 < t_1 < t_2 < \ldots < t_n$ the increments $Q(t_2) - Q(t_1),\ Q(t_3) - Q(t_2),\ \ldots,\ Q(t_n) - Q(t_{n-1})$ are independent.*

3. For $0 \leq s < t < \infty$ the increment $Q(t) - Q(s)$ has a Poisson distribution with parameter λ, i.e., the distribution of the increments is given by

$$P([Q(t) - Q(s)] = k) = \frac{\lambda^k (t-s)^k}{k!} e^{-\lambda(t-s)}$$

for $k \in N$.

The Poisson process is a continuous-time process with discrete realizations, because the state space contains only discrete numbers. The realizations are always positive by definition of N.

First, the probability of at least one event happening in a time period of duration Δt is given by

$$P(Q(t + \Delta t) - Q(t) \neq 0) = \lambda \Delta t + o(\Delta t^2)$$

with $\lambda > 0$ and $\Delta t \to 0$. The result is obtained by expanding the Taylor series of the Poisson distribution to the first order around $\Delta t \simeq 0$. Note that the probability of an event happening during Δt is proportional to the time period of duration. Second, the probability of two or more events happening during Δt is of order $o(\Delta t^2)$, therefore making this probability extremely small. Essentially, it says that the simultaneous occurrence of more than one event during a small Δt is almost zero.

Let $Q(t + \Delta t) - Q(t)$ be a Poisson process as defined above with the parameter λ. We formally define the differential $dQ(t)$ to be the limit

$$dQ(t) = \lim_{\Delta t \to dt} (Q(t + \Delta t) - Q(t)) \,.$$

From the definition of the Poisson process, it follows that $dQ(t)$ has the following properties:

1. $dQ(t) \quad = 0$ with probability $1 - \lambda dt$
2. $|dQ|(t) = 1$ with probability λdt .

Notes and Comments

Textbooks including the subject of stochastic processes are [7], [26], [3], [24], [30], [16], and [11].

3

Stochastic Differential Equations

Do not worry about your problems with mathematics,
I assure you mine are far greater.
Albert Einstein

Why should I refuse a good dinner simply because
I do not understand the digestive processes involved?
(Reply when criticised for his daring use of operators
before they could be justified formally.)
Oliver Heaviside

3.1 Introduction

In order to illustrate stochastic differential equations (SDEs), we first have a look at an ordinary differential equation (ODE). An ODE

$$\frac{dx(t)}{dt} = f(t, x) \tag{3.1}$$

may be viewed as a degenerated form of an SDE, as yet undefined, in the absence of randomness. We can write (3.1) in the symbolic differential form

$$dx(t) = f(t, x)dt$$

or, more accurately, as an integral equation

$$x(t) = x_0 + \int_0^t f(s, x(s))\, ds\,, \tag{3.2}$$

where $x(t) = x(t, x_0, t_0)$ is the solution satisfying the given initial condition $x(t_0) = x_0$. The idea of an ODE can be augmented to an SDE by adding noise to the system under consideration. For example, consider the following system

$$\frac{dx(t)}{dt} = a(t)x(t)\,, \quad x(0) = x_0\,, \tag{3.3}$$

where $a(t)$ is not a deterministic parameter. Rather, it is subjected to some random effects, so that we have $a(t) = f(t) + h(t)\xi(t)$. The uncertainty is

represented by the stochastic process $\xi(t)$. The differential equation (3.3) can now be rewritten as[1]

$$\frac{dX(t)}{dt} = f(t)X(t) + h(t)X(t)\xi(t)\,, \tag{3.4}$$

where $\xi(t)$ is a white noise process. By writing (3.4) in the differential form and using the substitution $dW(t) = \xi(t)dt$ (where $dW(t)$ is the differential form of a standard Brownian motion $W(t)$), we get the following SDE

$$dX(t) = f(t)X(t)dt + h(t)X(t)dW(t)\,.$$

In general, an SDE is given by

$$dX(t,\omega) = f(t, X(t,\omega))dt + g(t, X(t,\omega))dW(t,\omega)\,, \tag{3.5}$$

where ω indicates that $X = X(t,\omega)$ is a random process defined on the appropriate probability space and has the deterministic initial condition $X(0,\omega) = X_0$. For further discussions, we assume that $f(t, X(t,\omega)) \in \mathbb{R}$, $g(t, X(t,\omega)) \in \mathbb{R}$, and $W(t,\omega) \in \mathbb{R}$. In analogy to (3.2) we can rewrite (3.5) as

$$X(t,\omega) = X_0 + \int_0^t f(s, X(s,\omega))\, ds + \int_0^t g(s, X(s,\omega))\, dW(s,\omega)\,. \tag{3.6}$$

At the moment, there is still the problem of understanding what the integral $\int_0^t g(s, X(s,\omega))dW(s,\omega)$ exactly means. The first integral in (3.6) is quite familiar, since it is the ordinary Riemann integral. In order to define stochastic integrals, an approach similar to Riemann integrals is taken. This is presented in the next section.

3.2 Stochastic Integration or Itô Integrals

3.2.1 Definition

For the stochastic integral $\int_0^T g(t,\omega)dW(t,\omega)$, we assume that $g(t,\omega)$ changes only at the times t_i $(i = 1, 2, 3, ..., N-1)$, with

$$0 = t_0 < t_1 < t_2 < \ldots < t_{N-1} < t_N = T.$$

We define the integral

$$S = \int_0^T g(t,\omega)dW(t,\omega)\,, \tag{3.7}$$

[1] With capital letters we denote random variables or stochastic processes, whereas lower case letters denote deterministic variables or processes.

as the limit of the "Riemann sums"

$$S_N(\omega) = \sum_{i=1}^{N} g(t_{i-1}, \omega)\Big(W(t_i, \omega) - W(t_{i-1}, \omega)\Big)\,, \tag{3.8}$$

for $N \to \infty$. Unfortunately, the limit $\lim_{N\to\infty} S_N(\omega)$ cannot be treated as in the deterministic case. In the case of stochastic calculus in the Itô sense, the convergence concept of interest the is mean-square or L^2 convergence. In contrast to Riemann integrals, it matters, at which point of the interval $[t_{i-1}, t_i]$ the value of $g(t, \omega)$ is approximated. For Riemann integrals, the limit converges to the same number regardless of the chosen approximation. This is not the case for stochastic integrals and therefore, the limit depends on the choice of the point at which the approximation is considered. For stochastic calculus in the Itô sense, $g(t, \omega)$ is approximated at t_{i-1}. This choice implies the stochastic properties of the Itô integral which are discussed later in this chapter. Because we choose t_{i-1} for approximating $g(t, \omega)$, we call this approximation non-anticipative.

Definition 3.1. *A random variable S is called the Itô integral of a stochastic process $g(t, \omega)$ with respect to the Brownian motion $W(t, \omega)$ on the interval $[0, T]$ if*

$$\lim_{N\to\infty} E\Big[\Big(S - \sum_{i=1}^{N} g(t_{i-1}, \omega)\Big(W(t_i, \omega) - (W(t_{i-1}, \omega)\Big)\Big] = 0\,, \tag{3.9}$$

for each sequence of partitions $(t_0, t_1, \ldots, t_N)$ of the interval $[0, T]$ such that $\max_i(t_i - t_{i-1}) \to 0$.

In this case, (3.7) is defined and can be computed. There are some natural conditions which guarantee that a stochastic variable as described above does indeed exist, so that the integral is well defined. These conditions are stated later in the text. For a more detailed treatment of this definition refer to [3, Chapter 4].

The limit in the above definition converges to the stochastic integral in the mean-square sense. Thus, the stochastic integral is a random variable, the samples of which depend on the individual realizations of the paths $W(., \omega)$. The realizations of the random processes are observable in a path-wise sense adapted to the Brownian motion: At time $t = 0$, the stochastic integral is deterministic with value 0; however, at time $t = T$, the complete realizations of W have been observed and have contributed to the random value of the integral in the corresponding way.

3.2.2 Examples

To give a first illustration of the definition, let us assume that the argument is a non-random variable, such that $g(t, \omega) = g(t)$. The simplest possible example

is $g(t) = c$ for all t. This is still a stochastic process, but a simple one. Taking the definition, we actually get

$$\begin{aligned}\int_0^T c\, dW(t,\omega) &= c \lim_{N\to\infty} \sum_{i=1}^{N} \Big(W(t_i,\omega) - W(t_{i-1},\omega)\Big) \\ &= c \lim_{N\to\infty} [(W(t_1,\omega) - W(t_0,\omega)) + (W(t_2,\omega) - W(t_1,\omega)) + \dots \\ &\quad + (W(t_N,\omega) - W(t_{N-1},\omega)) \\ &= c\,(W(T,\omega) - W(0,\omega))\,,\end{aligned}$$

where $W(T,\omega)$ and $W(0,\omega)$ are standard Gaussian random variables. This makes sense, because it agrees with our intuition from standard calculus. To simplify things a bit, we now use the definition that the Brownian motion starts from zero, i.e., $W(0,\omega) = 0$. The last result becomes

$$\int_0^T c\, dW(t,\omega) = c\, W(T,\omega)\,.$$

In the next example, we allow that $g(t,\omega)$ itself is a random function. The random function can be approximated by a random step function

$$g(t,\omega) = g(t_{i-1},\omega)\,, \quad t \in [t_{i-1}, t_i]\,.$$

An illustrative example is $g(t,\omega) = W(t,\omega)$, so that the integrand itself is a random variable. The following algebraic identity is needed for the next calculation, $y(x-y) = yx - y^2 + \frac{1}{2}x^2 - \frac{1}{2}x^2 = \frac{1}{2}x^2 - \frac{1}{2}y^2 - \frac{1}{2}(x-y)^2$. By applying the definition for Itô integrals we obtain

$$\begin{aligned}\int_0^T W(t,\omega)\, dW(t,\omega) &= \lim_{N\to\infty} \sum_{i=1}^{N} W(t_{i-1},\omega)\Big(W(t_i,\omega) - W(t_{i-1},\omega)\Big) \\ &= \lim_{N\to\infty} \Big[\frac{1}{2}\sum_{i=1}^{N} \Big(W^2(t_i,\omega) - W^2(t_{i-1},\omega)\Big) \\ &\quad - \frac{1}{2}\sum_{i=1}^{N} \Big(W(t_i,\omega) - W(t_{i-1},\omega)\Big)^2\Big] \\ &= -\frac{1}{2} \lim_{N\to\infty} \sum_{i=1}^{N} \Big(W(t_i,\omega) - W(t_{i-1},\omega)\Big)^2 \\ &\quad + \frac{1}{2} W^2(T,\omega)\,.\end{aligned}$$

Due to the sum-of-squares property of the normalized Brownian motion W, the first term is deterministic, and we finally get

$$\int_0^T W(t,\omega)\, dW(t,\omega) = \frac{1}{2} W^2(T,\omega) - \frac{1}{2} T\,. \tag{3.10}$$

This is in contrast to our intuition from standard calculus. In the case of a deterministic integral $\int_0^T x(t)dx(t) = \frac{1}{2}x^2(t)$, whereas the Itô integral differs by the term $-\frac{1}{2}T$. — This example shows that the rules of differentiation (in particular the chain rule) and integration need to be re-formulated in the stochastic calculus.

3.2.3 Properties of Itô Integrals

We now state some important properties of Itô integrals. First, we compute the mean and the variance of the stochastic integral. For the mean, we obtain:

$$E[\int_0^T g(t,\omega)\, dW(t,\omega)] = 0\,.$$

Proof:

$$\begin{aligned} E[\int_0^T g(t,\omega)dW(t,\omega)] &= E[\lim_{N\to\infty}\sum_{i=1}^{N} g(t_{i-1},\omega)\Big(W(t_i,\omega) - W(t_{i-1},\omega)\Big)] \\ &= \lim_{N\to\infty}\sum_{i=1}^{N} E[g(t_{i-1},\omega)]\, E[\Big(W(t_i,\omega) - W(t_{i-1},\omega)\Big)] \\ &= 0\,. \end{aligned}$$

The expectation of stochastic integrals is zero. This is what we would expect anyway.

Next, the variance can be computed in a similar way and we obtain

$$\mathrm{Var}\Big[\int_0^T g(t,\omega)\, dW(t,\omega)\Big] = \int_0^T E[g^2(t,\omega)]\, dt\,.$$

Proof:

$$\begin{aligned} \mathrm{Var}\Big[\int_0^T g(t,\omega)\, dW(t,\omega)\Big] &= E\Big[(\int_0^T g(t,\omega)\, dW(t,\omega))^2\Big] \\ &= E\Big[\Big(\lim_{N\to\infty}\sum_{i=1}^{N} g(t_{i-1},\omega)\Big(W(t_i,\omega) - W(t_{i-1},\omega)\Big)\Big)^2\Big] \\ &= \lim_{N\to\infty}\sum_{i=1}^{N}\sum_{j=1}^{N} E[g(t_{i-1},\omega)g(t_{j-1},\omega) \\ &\quad (W(t_i,\omega) - W(t_{i-1},\omega))(W(t_j,\omega) - W(t_{j-1},\omega))] \\ &= \lim_{N\to\infty}\sum_{i=1}^{N} E[g^2(t_{i-1},\omega)]\, E[\Big(W(t_i,\omega) - W(t_{i-1},\omega)\Big)^2] \end{aligned}$$

$$= \lim_{N\to\infty} \sum_{i=1}^{N} E[g^2(t_{i-1},\omega)]\,(t_i - t_{i-1})$$
$$= \int_0^T E[g^2(t,\omega)]\,dt\,.$$

The computation of the variance yields another important property:

$$E\Big[\Big(\int_0^T g(t,\omega)dW(t,\omega)\Big)^2\Big] = \int_0^T E[g^2(t,\omega)]dt\,.$$

The condition that stochastic integrals are well-defined and solvable is briefly discussed here: More generally, one may ask which processes (or functions) $g(t,\omega)$ lead to well-defined results. The natural requirement is to impose on $g(t,\omega)$ that it does not depend on future values of the Brownian motion $W(t,\omega)$. When the stochastic process is interpreted as a stock price process, knowing tomorrow's asset price would make life on the stock market totally different from what it is today. Mathematically speaking, $g(t,\omega)$ should be adapted to the Brownian motion. This means that at any time t, the stochastic variable $g(t,\omega)$ depends only on the stochastic variable $\{W(t-h,\omega) \mid h \geq 0\}$ and possibly on other variables which are independent of the Brownian motion. For this reason, the integrand of the Itô integral is approximated by $g(t_{i-1},\omega)$. This is the non-anticipating approximation. Another requirement is that, roughly speaking, $g(t,\omega)$ should only attain large values with low probability, i.e.,

$$\int_0^T E[g^2(t,\omega)]dt < \infty\,.$$

Otherwise the variance would be without bounds. Of course, the Itô integral is linear,i.e.,

$$\int_0^T [a_1\, g_1(t,\omega) + a_2\, g_2(t,\omega)]dW(t,\omega)$$
$$= a_1 \int_0^T g_1(t,\omega)dW(t,\omega) + a_2 \int_0^T g_2(t,\omega)dW(t,\omega)$$

for any real numbers a_1, a_2 and any functions $g_1(t,\omega)$, $g_2(t,\omega)$. In the further discussion of this text, we shall omit the ω argument for convenience.

3.2.4 Stochastic Integrals for Poisson Processes

Let us consider the case that the source of uncertainty of an SDE is a Poisson process rather than a Brownian motion. In analogy to (3.5), we define an SDE, driven by a Poisson process as follows:

$$dX(t) = f(t,X(t))dt + h(t,X(t))A(t)dQ(t)\,, \tag{3.11}$$

where $dQ(t)$ is the differential form of the Poisson process and $X(0) = x_0$ is the initial condition. Furthermore, $dQ(t)$ is multiplied by $h(t, X(t), A(t))$, where $A(t)$ is an identically and independently distributed (i.i.d.) random variable which allows the "jump amplitude" to be an independent stochastic variable. Similarly to $g(t, X(t))$ for an SDE driven by Brownian motion, $h(t, X(t), A(t))$ allows us to model complicated non-linear stochastic environments. As before, we write (3.11) as the solution for $X(t)$ as follows:

$$X(t) = X_0 + \int_0^t f(s, X(s))ds + \sum_{i=1}^{N(t)} h(s_i, X(s_i))A(s_i)\,, \tag{3.12}$$

where $s_i \in [0, t]$ denote the points in time when a "jump" has occurred, $A(s_i)$ denotes the i-th jump amplitude drawn from the distribution $p(A)$, and $N(t)$ denotes the number of jumps that occurred between $[0, t]$. Since the Poisson process has discrete realizations, the stochastic integral degenerates to a finite sum. The term $\sum_{i=1}^{N(t)} h(s_i, X(s_i))A_i(s)$ adds up all Poisson events evaluated at the respective jump time s_i and with the jump amplitude A_i.

The most general SDE is one where a Brownian motion and a Poisson process are driving the uncertainty:

$$dX(t) = f(t, X(t))dt + g(t, X(t))dW(t) + h(t, X(t))A(t)dQ(t)\,,$$

where $W(t)$ is a standard Brownian motion and $Q(t)$ is a standard Poisson process.

3.3 Stochastic Differentials and Itô Calculus

3.3.1 The Scalar Case

As mentioned before, the rules of classical calculus are not valid for stochastic integrals and differential equations. In this section, we shall derive, heuristically, the famous *Itô formula*. It is the equivalent to the chain rule in classical calculus. The problem can be stated as follows:

Given a stochastic differential equation

$$dX(t) = f(t, X(t))dt + g(t, X(t))dW(t)\,, \tag{3.13}$$

and another process $Y(t)$ which is a function of $X(t)$,

$$Y(t) = \phi(t, X(t))\,,$$

where the function $\phi(t, X(t))$ is continuously differentiable in t and twice continuously differentiable in X, find the stochastic differential equation for the process $Y(t)$:

$$dY(t) = \tilde{f}(t, X(t))dt + \tilde{g}(t, X(t))dW(t)\,.$$

In the case when we assume that $g(t, X(t)) = 0$, we know the result: the chain rule for standard calculus. The result is given by

$$dy(t) = (\phi_t(t,x) + \phi_x(t,x)f(t,x))dt\,.$$

To arrive at a similar result in the case when $g(t, X(t)) \neq 0$, we reason as follows: The Taylor expansion of $\phi(t, X(t))$ yields

$$\begin{aligned} dY(t) = \phi_t(t,X)dt + \phi_x(t,X)dX(t) + \frac{1}{2}\phi_{tt}(t,X)dt^2 \\ +\frac{1}{2}\phi_{xx}(t,X)(dX(t))^2 + \phi_{xt}(t,X)dX(t)dt + \text{higher order terms}\,. \end{aligned}$$

Using the expression (3.13) for $dX(t)$ we arrive at

$$\begin{aligned} dY(t) = \phi_t(t,X)dt + \phi_x(t,X)[f(t,X(t))dt + g(t,X(t))dW(t)] \\ +\frac{1}{2}\phi_{tt}(t,X)dt^2 + \frac{1}{2}\phi_{xx}(t,X)[f(t,X(t))dt + g(t,X(t))dW(t)]^2 \\ +\phi_{xt}(t,X)[f(t,X(t))dt + g(t,X(t))dW(t)]dt + \text{higher order terms}\,. \end{aligned}$$

The higher-order differentials tend to be small compared to the first order terms dt and dW and thus $dt^2 \to 0$ and $dtdW(t) \to 0$. But, due to the sum-of-squares property of the Brownian motion W, we have $dW^2(t) = dt$. Thus, omitting higher order terms, we arrive at

$$\begin{aligned} dY(t) = \phi_t(t,X)dt + \phi_x(t,X)[f(t,X(t))dt + g(t,X(t))dW(t)] \\ +\frac{1}{2}\phi_{xx}(t,X)g^2(t,X(t))dt\,. \end{aligned}$$

Reordering the terms yields the scalar version of Itô's Lemma:

$$\begin{aligned} dY(t) = [\phi_t(t,X) + \phi_x(t,X)f(t,X(t)) + \frac{1}{2}\phi_{xx}(t,X)g^2(t,X(t))]dt \\ +\phi_x(t,X)g(t,X(t))dW(t)\,. \end{aligned}$$

Theorem 3.2. Itô's Lemma
Let $\phi(t, X(t))$ be a suitably differentiable function. For the random process defined by the stochastic differential equation

$$dX(t) = f(t,X(t))dt + g(t,X(t))dW(t)\,,$$

the transformed stochastic process $Y(t, X(t)) = \phi(t, X(t))$ satisfies the stochastic differential equation

$$dY(t) = \tilde{f}(t,X(t))dt + \tilde{g}(t,X(t))dW(t)\,,$$

where

$$\begin{aligned} \tilde{f}(t,X(t)) &= \phi_t(t,X) + \phi_x(t,X)f(t,X(t)) + \frac{1}{2}\phi_{xx}(t,X)g^2(t,X(t)) \\ \tilde{g}(t,X(t)) &= \phi_x(t,X)g(t,X(t))\,. \end{aligned}$$

The term $\frac{1}{2}\phi_{xx}(t,X)g^2(t,X(t))$ which appears in the expression for $\tilde{f}(t,X(t))$ is sometimes called the Itô correction term since it is absent in the deterministic case. Note that this term vanishes if ϕ is linear.

Sometimes, it is convenient to write the SDE of the process $Y(t)$ as

$$dY(t) = \phi_t(t,X)dt + \phi_x(t,X)dX + \frac{1}{2}\phi_{xx}(t,X)dX^2 \tag{3.14}$$

where dX^2 is computed according to the rules $dt^2 = 0$, $dt \cdot dW = 0$, and $dW^2 = dt$.

Example 3.3. $dX(t) = dW(t)$; $\phi(t,X(t)) = X^2(t)$. — From the SDE, we conclude that $X(t) = W(t)$. The relevant partial derivatives are: $\frac{\partial\phi(t,X)}{\partial X} = 2X$, $\frac{\partial^2\phi(t,X)}{\partial X^2} = 2$, and $\frac{\partial\phi(t,X)}{\partial t} = 0$. Itô's formula yields

$$d(W^2(t)) = dt + 2W(t)dW(t)\,.$$

Rewriting the equation in integral form and using $W(0) = 0$, we get

$$W^2(t) = t + 2\int_0^t W(t)dW(t)$$

or

$$\int_0^t W(t)dW(t) = \frac{1}{2}W^2(t) - \frac{1}{2}t\ .$$

This is exactly the result we obtained by solving the stochastic integral in (3.10). From this result, we conclude that our intuition based on standard calculus does not work. This is confirmed by the Itô correction term.

3.3.2 The Vector Case

All of the above was for a situation in which the process $X(t)$ is scalar. We now allow that the process $X(t)$ is in $\mathbb{R}^n$. We let $W(t)$ be an m-dimensional standard Brownian motion. This means that $W(t)$ takes values in $\mathbb{R}^m$ and its components processes $W_i(t), (i = 1, 2, ..., m)$ are independent scalar Brownian motions. To make the dimensions fit, we have $f(t,X(t)) \in \mathbb{R}^n$ and $g(t,X(t)) \in \mathbb{R}^{n\times m}$. Consider a scalar process $Y(t)$ defined by $Y(t) = \phi(t,X(t))$, where $\phi(t,X)$ is a scalar function which is continuously differentiable with respect to t and twice continuously differentiable with respect to X. The Itô formula can be written in vector notation as follows:

$$dY(t) = \tilde{f}(t,X(t))dt + \tilde{g}(t,X(t))dW(t)$$

with

$$\begin{aligned}\tilde{f}(t,X(t)) &= \phi_t(t,X(t)) + \phi_x(t,X(t))f(t,X(t)) \\ &\quad + \frac{1}{2}\mathrm{tr}\Big(\phi_{xx}(t,X(t))g(t,X(t))g^T(t,X(t))\Big) \\ &= \phi_t(t,X(t)) + \phi_x(t,X(t))f(t,X(t)) \\ &\quad + \frac{1}{2}\mathrm{tr}\Big(g^T(t,X(t))\phi_{xx}(t,X(t))g(t,X(t))\Big)\end{aligned}$$

and

$$\tilde{g}(t,X(t)) = \phi_x(t,X(t))g(t,X(t)) \ ,$$

where "tr" denotes the trace operator.

3.3.3 Examples

In this section, we show how to use Itô's formula. Most of the examples are classical examples which can be found in numerous textbooks.

Example 3.4. Consider the following stochastic differential equation:

$$dS(t) = \mu\, S(t)dt + \sigma\, S(t)dW(t)\,.$$

The process $S(.)$ is called geometric Brownian motion. We want to find the stochastic differential equation for the process Y related to S as follows:

$$Y(t) = \phi(t,S(t)) = \log(S(t))\,. \tag{3.15}$$

The relevant partial derivatives are $\frac{\partial\phi(t,S)}{\partial S} = \frac{1}{S}$, $\frac{\partial^2\phi(t,S)}{\partial S^2} = -\frac{1}{S^2}$, and $\frac{\partial\phi(t,S)}{\partial t} = 0$. Therefore, according to Itô

$$\begin{aligned}dY(t) &= \Big(\frac{\partial\phi(t,S)}{\partial t} + \frac{\partial\phi(t,S)}{\partial S}\mu S(t) + \frac{1}{2}\frac{\partial^2\phi(t,S)}{\partial S^2}\sigma^2S^2(t)\Big)dt \\ &\quad + \Big(\frac{\partial\phi(t,S)}{\partial S}\sigma S(t)\Big)dW(t) \\ dY(t) &= (\mu - \frac{1}{2}\sigma^2)dt + \sigma dW(t)\,.\end{aligned} \tag{3.16}$$

Obviously, $Y(.)$ is a non-standard Brownian motion with

$$\begin{aligned}E[Y(t)] &= (\mu - \frac{1}{2}\sigma^2)t \quad \text{and} \\ \mathrm{Var}[Y(t)] &= \sigma^2 t\,.\end{aligned}$$

So, for the sample path of $Y(.)$, we have the following closed-form solution:

$$Y(t) = Y_0 + (\mu - \frac{1}{2}\sigma^2)t + \sigma W(t)\,.$$

Hence, for the sample path of the geometric Brownian motion $S(.) = e^{Y(.)}$, we have the closed-form solution

$$S(t) = e^{Y_0} e^{(\mu - \frac{1}{2}\sigma^2)t + \sigma W(t)} \,. \tag{3.17}$$

The geometric Brownian motion is a popular stock price model and (3.17) gives the exact solution for this price process.

Example 3.5. Suppose that two processes $X_1(t)$ and $X_2(t)$ are given by the *coupled* SDEs

$$\begin{bmatrix} dX_1(t) \\ dX_2(t) \end{bmatrix} = \begin{bmatrix} f_1(t, X(t)) \\ f_2(t, X(t)) \end{bmatrix} dt + \begin{bmatrix} g_1(t, X(t)) & 0 \\ 0 & g_2(t, X(t)) \end{bmatrix} \begin{bmatrix} dW_1(t) \\ dW_2(t) \end{bmatrix},$$

where $W_1(t)$ and $W_2(t)$ are two independent Brownian motions. Let us now compute a stochastic differential for the product $Y = X_1 X_2$. Thus, $\phi(t, X_1, X_2) = X_1 X_2$. An interpretation of ϕ could be that X_1 describes the evolution of an American stock in US \$, whereas X_2 describes the evolution of the exchange rate CHF/US \$ and thus ϕ describes the evolution of the price of the American stock measured in CHF. Let us now apply the multivariate form of the Itô calculus. The partial derivatives are

$$\begin{aligned} \phi_t(t, X_1, X_2) &= 0 \\ \phi_{x_1}(t, X_1, X_2) &= X_2 \\ \phi_{x_2}(t, X_1, X_2) &= X_1 \\ \phi_{x_1 x_1}(t, X_1, X_2) = \phi_{x_2 x_2}(t, X_1, X_2) &= 0 \\ \phi_{x_1 x_2}(t, X_1, X_2) = \phi_{x_2 x_1}(t, X_1, X_2) &= 1 \,. \end{aligned}$$

Having computed the elements of the Jacobian and the Hessian, we can calculate the SDE for $Y(t)$:

$$\begin{aligned} \tilde{f}(t, X(t)) &= \phi_t(t, X) + \phi_x(t, X) f(t, X(t)) \\ &\quad + \frac{1}{2} \mathrm{tr}\Big(g^T(t, X(t)) \phi_{xx}(t, X) g(t, X(t)) \Big) \\ &= X_2(t) f_1(t, X(t)) + X_1(t) f_2(t, X(t)) \end{aligned}$$

$$\begin{aligned} \tilde{g}(t, X(t)) dW(t) &= [\phi_x(t, X) g(t, X(t))] dW(t) \\ &= X_2(t) g_1(t, X(t)) dW_1(t) + X_1(t) g_2(t, X(t)) dW_2(t) \end{aligned}$$

$$\begin{aligned} dY(t) &= [X_2(t) f_1(t, X(t)) + X_1(t) f_2(t, X(t))] dt \\ &\quad + X_2(t) g_1(t, X(t)) dW_1(t) + X_1(t) g_2(t, X(t)) dW_2(t) \end{aligned}$$

or, with the with the almost suspiciously simple final result,

$$dY(t) = dX_1(t) X_2(t) + X_1(t) dX_2(t) \,.$$

The SDE for $Y(t)$ shows that the investor faces two sources of uncertainties: the first from the uncertainty of the American stock and the second from the uncertainty of the exchange rate. Mathematically, the risk is represented by the two Brownian motions, which drive the uncertainty of the SDE for $Y(t)$.

Example 3.6. Suppose that two processes $X_1(t)$ and $X_2(t)$ are given by the *uncoupled* SDEs

$$\begin{bmatrix} dX_1(t) \\ dX_2(t) \end{bmatrix} = \begin{bmatrix} f_1(t, X_1(t)) \\ f_2(t, X_2(t)) \end{bmatrix} dt + \begin{bmatrix} g_1(t, X_1(t)) \\ g_2(t, X_2(t)) \end{bmatrix} dW(t)\,,$$

Let us compute a stochastic differential for the product $Y = X_1 X_2$. Thus, $\phi(t, X_1, X_2) = X_1 X_2$. Let us apply the multivariate form of the Itô calculus. The partial derivatives are

$$\begin{aligned} \phi_t(t, X_1, X_2) &= 0 \\ \phi_{x_1}(t, X_1, X_2) &= X_2 \\ \phi_{x_2}(t, X_1, X_2) &= X_1 \\ \phi_{x_1 x_1}(t, X_1, X_2) = \phi_{x_2 x_2}(t, X_1, X_2) &= 0 \\ \phi_{x_1 x_2}(t, X_1, X_2) = \phi_{x_2 x_1}(t, X_1, X_2) &= 1\,. \end{aligned}$$

Having computed the elements of the Jacobian and the Hessian, we can calculate the SDE for $Y(t)$:

$$\begin{aligned} \tilde{f}(t, X(t)) &= \phi_t(t, X) + \phi_x(t, X) f(t, X(t)) \\ &\quad + \frac{1}{2} \mathrm{tr}\Big(g^T(t, X(t)) \phi_{xx}(t, X) g(t, X(t)) \Big) \\ &= X_2(t) f_1(t, X_1(t)) + X_1(t) f_2(t, X_2(t)) \\ &\quad + g_1(t, X_1(t)) g_2(t, X_2(t)) \end{aligned}$$

$$\begin{aligned} \tilde{g}(t, X(t)) dW(t) &= \phi_x(t, X) g(t, X(t)) dW(t) \\ &= [X_2(t) g_1(t, X_1(t)) + X_1(t) g_2(t, X_2(t))] dW(t) \end{aligned}$$

$$\begin{aligned} dY(t) &= [X_2(t) f_1(t, X_1(t)) + X_1(t) f_2(t, X_2(t)) \\ &\quad + g_1(X_1(t)) g_2(X_2(t))] dt \\ &\quad + [X_2(t) g_1(t, X_1(t)) + X_1(t) g_2(t, X_2(t))] dW(t) \end{aligned}$$

or, with the surprisingly simple final result,

$$dY(t) = dX_1(t) X_2(t) + X_1(t) dX_2(t) + g_1(t, X_1(t)) g_2(t, X_2(t)) dt\,.$$

3.3.4 Itô Calculus for Poisson Processes

We briefly state the differentiation rules for scalar SDEs driven by a Brownian motion and a Poisson process. The dynamics of $X(t)$ can be described as a linear superposition of an SDE driven by a Brownian motion and an SDE driven by a Poisson process.

Theorem 3.7. Itô calculus for Poisson processes
Let $\phi(t, X(t))$ be a twice continuously differentiable function. For the stochastic differential equation

$$dX(t) = f(t, X(t))dt + g(t, X(t))dW(t) + h(t, X(t), A(t))dQ(t)\,,$$

the transformation $Y(t, X(t)) = \phi(t, X(t))$ *satisfies the stochastic differential equation*

$$dY(t) = \tilde{f}(t, X(t))dt + \tilde{g}(t, X(t))dW(t) + \tilde{h}(t, X(t), A(t))dQ(t),$$

with

$$\begin{aligned}
\tilde{f}(t, X(t)) &= \phi_t(t, X) + \phi_x(t, X)f(t, X(t)) + \frac{1}{2}\phi_{xx}(t, X)g^2(t, X(t)) \\
\tilde{g}(t, X(t)) &= \phi_x(t, X)g(t, X(t)) \\
\tilde{h}(t, X(t), A(t)) &= \phi(t, X(t) + h(t, X(t), A(t))\,) - \phi(t, X(t))\,.
\end{aligned}$$

In the case where $h(t, X(t), A(t)) = 0$, we recover the Itô formula in the scalar case. The transformation in front of the $dQ(t)$ operator is the transformation before and after a jump has occurred.

Example 3.8. The example is a geometric Brownian motion with Poisson process which allows for downward jumps of the stock price process under consideration

$$dS(t) = \mu S(t)dt + \sigma S(t)dW(t) + (e^{-A(t)} - 1)S(t)dQ(t)\,,$$

where μ and σ are constants, $A(t) > 0$ is the magnitude of the stochastic jump, and the Poisson process $Q(t)$ has the jump frequency λ. Hence $(e^{-A(t)} - 1)$ represents a downward jump which leaves $S(t)$ positive. We assume that the jump magnitude is determined by draws from an exponential distribution with density

$$p(A) = \frac{1}{\eta}e^{-\frac{A}{\eta}}\,.$$

We want to compute the SDE of the transformation $Y(t) = \phi(t, X(t)) = \log(S(t))$. The relevant partial derivatives are $\frac{\partial\phi(t,S)}{\partial S} = \frac{1}{S}$, $\frac{\partial^2\phi(t,S)}{\partial S^2} = -\frac{1}{S^2}$, and $\frac{\partial\phi(t,S)}{\partial t} = 0$. Applying the extended Itô formula we get

$$\begin{aligned}
dY(t) = &\Big(\frac{\partial\phi(t,S)}{\partial t} + \frac{\partial\phi(t,S)}{\partial S}\mu S(t) + \frac{1}{2}\frac{\partial^2\phi(t,S)}{\partial S^2}\sigma^2 S^2(t)\Big)dt \\
&+\frac{\partial\phi(t,S)}{\partial S}\sigma S(t)dW(t) \\
&+[\log(S(t) + (e^{-A(t)} - 1)S(t)) - \log(S(t)]dQ(t) \\
dY(t) = &(\mu - \frac{1}{2}\sigma^2)dt + \sigma dW(t) - A(t)dQ(t)\,.
\end{aligned}$$

3.4 Stochastic Differential Equations

We now possess the main tools to deal with stochastic differential equations, namely the Itô calculus and the stochastic integration. We classify SDEs

into two large groups, linear SDEs and non-linear SDEs. Furthermore, we distinguish between scalar linear and vector-valued linear SDEs. For both classes, examples from economics, finance, and engineering are given. Commonly, $f(t, X(t))$ is called *drift* and $g(t, X(t))$ is called *diffusion*.

3.4.1 Linear Scalar SDEs

In the case of linear SDEs, we can derive exact solutions and we are able to compute moments.

Definition 3.9. *A stochastic differential equation*

$$dX(t) = f(t, X(t))dt + g(t, X(t))dW(t)$$

with the initial condition $X(0) = x_0$ for a one-dimensional stochastic process $X(t)$ is called a linear (scalar) SDE if and only if the functions $f(t, X(t))$ and $g(t, X(t))$ are affine functions of $X(t) \in \mathbb{R}$ and thus

$$\begin{aligned} f(t, X(t)) &= A(t)X(t) + a(t)\,, \\ g(t, X(t)) &= [B_1(t)X(t) + b_1(t), \cdots, B_m(t)X(t) + b_m(t)]\,, \end{aligned}$$

where $A(t), a(t) \in \mathbb{R}$, $W(t) \in \mathbb{R}^m$ is an m-dimensional Brownian motion, and $B_i(t), b_i(t) \in \mathbb{R}$, $i = 1, \cdots, m$. Hence, $f(t, X(t)) \in \mathbb{R}$ and $g(t, X(t)) \in \mathbb{R}^{1\times m}$.

Alternatively, the scalar linear SDE can be written in the form

$$dX(t) = (A(t)X(t) + a(t))dt + \sum_{i=1}^{m}(B_i(t)X(t) + b_i(t))dW_i(t)\ .^2 \tag{3.18}$$

This linear SDE has the following solution:

$$\begin{aligned} X(t) = \Phi(t)\Big(x_0 + \int_0^t \Phi^{-1}(s)\Big[a(s) - \sum_{i=1}^{m} B_i(s)b_i(s)\Big]ds \\ + \sum_{i=1}^{m}\int_0^t \Phi^{-1}(s)b_i(s)\, dW_i(s)\Big)\,, \end{aligned} \tag{3.19}$$

where the fundamental matrix $\Phi(t)$ is given by

$$\Phi(t) = \exp\Big(\int_0^t \Big[A(s) - \sum_{i=1}^{m}\frac{B_i^2(s)}{2}\Big]ds + \sum_{i=1}^{m}\int_0^t B_i(s)dW_i(s)\Big)\,. \tag{3.20}$$

[2] A common extension of this equation is the following form of a controlled stochastic differential equation as given by

$$dX(t) = (A(t)X(t)+C(t)u(t)+a(t))dt + \sum_{i=1}^{m}(B_i(t)X(t)+D_i(t)u(t)+b_i(t))dW_i\,,$$

with the control $u(t) \in \mathbb{R}^k$, and $C(t) \in \mathbb{R}^{1\times k}$, and $D_i(t) \in \mathbb{R}^{1\times k}$.

This is the solution of the following SDE:

$$d\Phi(t) = A(t)\Phi(t)dt + \sum_{i=1}^{m} B_i(t)\Phi(t)dW_i(t)\,,$$

with the initial condition $\Phi(0) = 1$. For the proof, the reader is referred to [3, Chapter 8].

Example 3.10. Let us assume that $W(t) \in \mathbb{R}$, $a(t) = 0$, $b(t) = 0$, $A(t) = A$, $B(t) = B$. We want to compute the solution of the SDE

$$\begin{aligned} dX(t) &= AX(t)dt + BX(t)dW(t) \\ X(0) &= x_0\,. \end{aligned}$$

We can solve it using (3.19) and (3.20). As we already know, the solution is:

$$X(t) = x_0 e^{(A-\frac{1}{2}B^2)t+BW(t)}\,.$$

For scalar linear SDEs, ordinary differential equations exist for computing their first and second moments of the random process $X(t)$. The expectation $m(t) = E[X(t)]$ and the second moment $P(t) = E[X^2(t)]$ for (3.18) can be calculated by solving the following system of ODEs:

$$\begin{aligned} &\dot{m}(t) = A(t)m(t) + a(t) \\ &m(0) = x_0 \\ &\dot{P}(t) = \Big(2A(t) + \sum_{i=1}^{m} B_i^2(t)\Big)P(t) + 2m(t)\Big(a(t) + \sum_{i=1}^{m} B_i(t)b_i(t)\Big) + \sum_{i=1}^{m} b_i^2(t) \\ &P(0) = x_0^2\,. \end{aligned}$$

The ODE for the expectation is derived by applying the expectation operator on both sides of (3.19). Using the rules for the expectation operator, we get

$$\begin{aligned} \mathrm{E}[dX(t)] &= \mathrm{E}[(A(t)X(t) + a(t))dt + \sum_{i=1}^{m}(B_i(t)X(t) + b_i(t))dW_i(t)\,] \\ \underbrace{\mathrm{E}[dX(t)]}_{dm(t)} &= (A(t)\underbrace{\mathrm{E}[X(t)]}_{=m(t)} + a(t))dt + \sum_{i=1}^{m}\mathrm{E}[(B_i(t)X(t) + b_i(t))]\underbrace{\mathrm{E}[dW_i(t)\,]}_{=0} \\ dm(t) &= (A(t)m(t) + a(t))dt\,. \end{aligned}$$

In order to compute the second moment, we need to derive the SDE for $Y(t) = X^2(t)$:

$$
\begin{aligned}
dY(t) &= \Big[2X(t)(A(t)X(t)+a(t)) + \sum_{i=1}^{m}\Big(B_i(t)X(t)+b_i(t)\Big)^2\Big]dt \\
&\quad +2X(t)\sum_{i=1}^{m}\Big(B_i(t)X(t)+b_i(t)\Big)dW_i(t) \\
&= \Big[2A(t)X^2(t)+2X(t)a(t)+\sum_{i=1}^{m}\Big(B_i^2(t)X^2(t)+2B_i(t)b_i(t)X(t) \\
&\quad +b_i^2(t)\Big)\Big]dt + 2X(t)\sum_{i=1}^{m}\Big(B_i(t)X(t)+b_i(t)\Big)dW_i(t)\ . \qquad (3.21)
\end{aligned}
$$

Furthermore, we apply the expectation operator to (3.21) and use $P(t) = \mathrm{E}[X^2(t)] = \mathrm{E}[Y(t)]$ and $m(t) = \mathrm{E}[X(t)]$. We obtain

$$
\begin{aligned}
\underbrace{\mathrm{E}[dY(t)]}_{=dP(t)} &= \Big[2A(t)\underbrace{\mathrm{E}[X^2(t)]}_{P(t)} + 2a(t)\underbrace{\mathrm{E}[X(t)]}_{=m(t)} + \sum_{i=1}^{m}\Big(B_i^2(t)\underbrace{\mathrm{E}[X^2(t)]}_{=P(t)} \\
&\quad +2B_i(t)b_i(t)\underbrace{\mathrm{E}[X(t)]}_{=m(t)} + b_i^2(t)\Big)\Big]dt \\
&\quad +\mathrm{E}\Big[2X(t)\sum_{i=1}^{m}\Big(B_i(t)X(t)+b_i(t)\Big)\Big]\underbrace{\mathrm{E}\Big[dW_i(t)\Big]}_{=0} \\
dP(t) &= \Big[2A(t)P(t)+2a(t)m(t) \\
&\quad +\sum_{i=1}^{m}\Big(B_i^2(t)P(t)+2B_i(t)b_i(t)m(t)+b_i^2(t)\Big)\Big]dt\,.
\end{aligned}
$$

In the special case where $B_i(t) = 0,\ i = 1, \ldots, m$, we can explicitly give the exact probability density function: The solution of the scalar linear SDE

$$
dX(t) = (A(t)X(t)+a(t))dt + \sum_{i=1}^{m} b_i(t)dW_i(t) \qquad (3.22)
$$

with the initial condition $X(0) = x_0$ is normally distributed:

$$
P(X(t)|x_0) \sim \mathcal{N}(m(t), V(t))
$$

with the mean $m(t)$ and the variance $V(t) = P(t) - m^2(t)$, which are the solutions of the following ODEs:

$$
\begin{aligned}
\dot{m}(t) &= A(t)m(t)+a(t) \\
m(0) &= x_0 \\
\dot{V}(t) &= 2A(t)V(t)+\sum_{i=1}^{m} b_i^2(t) \\
V(0) &= 0\,.
\end{aligned}
$$

3.4.2 Popular Scalar Linear Models

There are some specific scalar linear SDEs which are found to be quite useful in practice. In this subsection, we discuss these cases and show what types of asset prices they could represent and how they could be useful. Examples of price processes can be found in [27] and [28].

- Brownian motion:
  The simplest case of stochastic differential equations is where the drift and the diffusion coefficients are independent of the information received over time

$$dS(t) = \mu dt + \sigma dW(t,\omega)\,, \quad S(0) = S_0\,.$$

  This model has been used to simulate commodity prices, such as metals or agricultural products. The mean is $\mathrm{E}[S(t)] = \mu t + S_0$ and the variance $\mathrm{Var}[S(t)] = \sigma^2 t$. $S(t)$ possesses a behavior of fluctuations around the straight line $S_0 + \mu t$. The process is normally distributed with the given mean and variance. The parameter σ is also called volatility. The volatility is used in many asset models as an indicator of the risk an investor is taking by buying a certain asset.

- Geometric Brownian motion:
  The standard model of stock prices is the geometric Brownian motion as given by

$$dS(t) = \mu S(t)dt + \sigma S(t)dW(t,\omega)\,, \quad S(0) = S_0\,.$$

  The mean is given by $\mathrm{E}[S(t)] = S_0 e^{\mu t}$ and its variance by $\mathrm{Var}[S(t)] = S_0^2 e^{2\mu t}(e^{\sigma^2 t} - 1)$. This model forms the starting point for the famous Black-Scholes formula for option pricing. The geometric Brownian motion has two main features which make it popular for stock price models: The first property is that $S(t) > 0$ for all $t \in [0,T]$ and the second is that all returns are in scale with the current price. This process has a log-normal probability density function. For $\mu > 0$, the moments of the geometric Brownian motion become infinite, and thus the model is unstable:

$$\lim_{t\to\infty} \mathrm{E}[S(t)] = \lim_{t\to\infty} S_0 e^{\mu t} \to \infty$$
$$\lim_{t\to\infty} \mathrm{Var}[S(t)] = \lim_{t\to\infty} S_0^2 e^{2\mu t}(e^{\sigma^2 t} - 1) \to \infty\,.$$

- Mean reverting process:
  Another very popular class of SDEs are mean reverting linear SDEs. This class of SDE is used to model for example short rates of the term structure of interest rates or electricity prices. The model is obtained by

$$dS(t) = \kappa[\mu - S(t)]dt + \sigma\, dW(t,\omega)\,, \quad S(0) = S_0\,. \tag{3.23}$$

A special case of this SDE where $\mu = 0$ is called *Ohrnstein-Uhlenbeck* process. Equation (3.23) models a process which naturally falls back to its equilibrium level of μ. The parameter $\kappa > 0$ governs how fast the process moves back to μ. When the price $S(t)$ is above μ, then $\kappa(\mu - S(t)) < 0$ and the probability that $S(t)$ decreases is high and when $S(t)$ is below μ, then $\kappa(\mu - S(t)) > 0$ and the probability is high that $S(t)$ increases. The expected price is $\mathrm{E}[S(t)] = \mu - (\mu - S_0)e^{-\kappa\, t}$ and the variance is

$$\mathrm{Var}[S(t)] = \frac{\sigma^2}{2\kappa}\Big(1 - e^{-2\kappa\, t}\Big)\,.$$

In the long run, the following (unconditional) approximations are valid

$$\lim_{t\to\infty} \mathrm{E}[S(t)] = \mu$$

and

$$\lim_{t\to\infty} \mathrm{Var}[S(t)] = \frac{\sigma^2}{2\kappa}\,.$$

This analysis shows that the process fluctuates around μ and has a variance of $\frac{\sigma^2}{2\kappa}$ which depends on the parameter κ: the higher κ, the lower the variance. This is obvious since the higher κ, the faster the process reverts back to its mean value. This process is a stationary process which is normally distributed. There are many variations of the mean reverting process. A popular extension is where the diffusion term is in scale with the current value, i.e., the geometric mean reverting process:

$$dS(t) = \kappa[\mu - S(t)]dt + \sigma S(t) dW(t,\omega)\,, \quad S(0) = S_0\,.$$

- Engineering model:
  In control engineering science, the most important (scalar) case is

$$dX(t) = (A(t)X(t) + C(t)u(t))\, dt + \sum_{i=1}^{m} b_i(t)\, dW_i\,. \tag{3.24}$$

In this equation, $X(t)$ is normally distributed because the Brownian motion is just multiplied by time-dependent factors. When we compute an optimal control law for this SDE, the deterministic optimal control law (ignoring the Brownian motion) and the stochastic optimal control law are the same. This feature is called *certainty equivalence*. For this reason, the stochastics are often ignored in control engineering.

3.4.3 Vector-Valued Linear SDEs

The logical extension of scalar SDEs is to allow $X(t) \in \mathbb{R}^n$ to be a vector. The rest of this section proceeds in a similar fashion as for scalar linear SDEs.

Definition 3.11. *A stochastic vector differential equation*

$$dX(t) = f(t, X(t))dt + g(t, X(t))dW(t)$$

with the initial condition $X(0) = x_0 \in \mathbb{R}^n$ *for an n-dimensional stochastic process* $X(t)$ *is called a linear SDE if the functions* $f(t, X(t)) \in \mathbb{R}^n$ *and* $g(t, X(t)) \in \mathbb{R}^{n\times m}$ *are affine functions of* $X(t)$ *and thus*

$$\begin{aligned} f(t, X(t)) &= A(t)X(t) + a(t)\,, \\ g(t, X(t)) &= [B_1(t)X(t) + b_1(t), \cdots, B_m(t)X(t) + b_m(t)]\,, \end{aligned}$$

where $A(t) \in \mathbb{R}^{n\times n}$, $a(t) \in \mathbb{R}^n$, $W(t) \in \mathbb{R}^m$ *is an m-dimensional Brownian motion, and* $B_i(t) \in \mathbb{R}^{n\times n}$, $b_i(t) \in \mathbb{R}^n$.

Alternatively, the vector-valued linear SDE can be written as

$$dX(t) = (A(t)X(t) + a(t))dt + \sum_{i=1}^{m}(B_i(t)X(t) + b_i(t))dW_i(t)\,. \tag{3.25}$$

A common extension of the above equation is the following form of a controlled stochastic differential equation as given by

$$\begin{aligned} dX(t) = &\left(A(t)X(t) + C(t)u(t) + a(t)\right)dt \\ &+ \sum_{i=1}^{m}\left(B_i(t)X(t) + D_i(t)u(t) + b_i(t)\right)dW_i\,, \end{aligned} \tag{3.26}$$

where $u(t) \in \mathbb{R}^k$, $C(t) \in \mathbb{R}^{n\times k}$, $D_i(t) \in \mathbb{R}^{n\times k}$. The linear SDE (3.25) has the following solution:

$$\begin{aligned} X(t) = \Phi(t)\Big(x_0 + \int_0^t \Phi^{-1}(s)\Big[a(s) - \sum_{i=1}^{m} B_i(s)b_i(s)\Big]ds \\ + \sum_{i=1}^{m}\int_0^t \Phi^{-1}(s)b_i(s)dW_i(s)\Big)\,, \end{aligned} \tag{3.27}$$

where the fundamental matrix $\Phi(t) \in \mathbb{R}^{n\times n}$ is the solution of the homogenous stochastic differential equation

$$d\Phi(t) = A(t)\Phi(t)dt + \sum_{i=1}^{m} B_i(t)\Phi(t)dW_i(t)\,, \tag{3.28}$$

with the initial condition $\Phi(0) = I \in \mathbb{R}^{n\times n}$. We now prove that (3.27) and (3.28) are solutions of (3.25). We rewrite (3.27) as

$$
\begin{aligned}
X(t) &= \Phi(t)\Big(x_0 + \int_0^t \Phi^{-1}(t)dY(t)\Big) \\
dY(t) &= \Big[a(t) - \sum_{i=1}^m B_i(t)b_i(t)\Big]dt + \sum_{i=1}^m b_i(t)dW_i(t) \\
X(t) &= \Phi(t)Z(t) \\
Z(t) &= x_0 + \int_0^t \Phi^{-1}(t)dY(t) \\
dZ(t) &= \Phi^{-1}(t)dY(t)\,.
\end{aligned}
$$

In order to prove that $X(t) = \Phi(t)Z(t)$ solves the SDE (3.25), we compute $dX(t)$ and show that this gives us back (3.25). Using Itô's formula, we get

$$
\begin{aligned}
dX(t) &= \Phi(t)dZ(t) + d\Phi(t)Z(t) + \sum_{i=1}^m B_i(t)\Phi(t)\Phi(t)^{-1}b_i(t)dt \\
&= dY(t) + A(t)\Phi(t)Z(t)dt + \sum_{i=1}^m B_i(t)\Phi(t)Z(t)dW_i(t) + \sum_{i=1}^m B_i(t)b_i(t)dt\,.
\end{aligned}
$$

Noting that $Z(t) = \Phi^{-1}(t)X(t)$ and using the SDE for $Y(t)$, we get

$$
\begin{aligned}
dX(t) &= dY(t) + A(t)\Phi(t)Z(t)dt + \sum_{i=1}^m B_i(t)\Phi(t)Z(t)dW_i(t) + \sum_{i=1}^m B_i(t)b_i(t)dt \\
&= \Big[a(t) - \sum_{i=1}^m B_i(t)b_i(t)\Big]dt + \sum_{i=1}^m b_i(t)dW_i(t) + A(t)X(t)dt \\
&\quad + \sum_{i=1}^m B_i(t)X(t)dW_i(t) + \sum_{i=1}^m B_i(t)b_i(t)dt \\
&= [a(t) + A(t)X(t)]dt + \sum_{i=1}^m (B_i(t)X(t) + b_i(t))dW_i(t)\,.
\end{aligned}
$$

This completes the proof.

For vector-valued linear models, there exist ordinary differential equations for computing the moments of the stochastic process $X(t)$. The expectation $m(t) = \mathrm{E}[X(t)] \in \mathbb{R}^n$ and the second moment matrix $P(t) = \mathrm{E}[X(t)X^T(t)] \in \mathbb{R}^{n\times n}$ can be computed as follows:

$$
\begin{aligned}
\dot{m}(t) &= A(t)m(t) + a(t) \\
m(0) &= x_0 \\
\dot{P}(t) &= A(t)P(t) + P(t)A^T(t) + a(t)m^T(t) + m(t)a^T(t) \\
&\quad + \sum_{i=1}^m \Big[B_i(t)P(t)B_i^T(t) + B_i(t)m(t)b_i^T(t) \\
&\qquad + b_i(t)m^T(t)B_i^T(t) + b_i(t)b_i(t)^T\Big] \\
P(0) &= x_0x_0^T\,.
\end{aligned}
$$

The covariance matrix for the system of linear SDEs is given by

$$V(t) = \mathrm{Var}\{x(t)\} = P(t) - m(t)m^T(t)\,.$$

The ease of calculating moments is one of the advantages of linear SDEs. Furthermore, due to the general solution theory, the solutions of many linear SDEs can explicitly be computed. In the special case where $B_i(t) = 0,\ i = 1, \ldots, m$, we can explicitly give the solution in form of a probability density function: The solution of the linear vector SDE

$$dX(t) = (A(t)X(t) + a(t))dt + \sum_{i=1}^{m} b_i(t)dW_i(t)$$

with the initial condition $X(0) = x_0 \in \mathbb{R}^n$ is normally distributed, i.e.,

$$P(X(t)|x_0) \sim \mathcal{N}(m(t), V(t))$$

with the expectation $m(t) \in \mathbb{R}^n$ and the covariance matrix $V(t) \in \mathbb{R}^{n\times n}$ which are the solutions of the following ODEs:

$$\begin{aligned}
\dot{m}(t) &= A(t)m(t) + a(t)\\
m(0) &= x_0\\
\dot{V}(t) &= A(t)V(t) + V(t)A^T(t) + \sum_{i=1}^{m} b_i b_i^T(t)\\
V(0) &= 0\,.
\end{aligned}$$

3.4.4 Popular Vector-Valued Linear Price Models

In this part, we show two popular multi-dimensional linear price models:

- Multi-dimensional geometric Brownian motion:
  The most popular stock price model is the geometric Brownian motion. In order to model two stock price processes, which are correlated and have the properties of one-dimensional geometric Brownian motions, we propose the following system of equations:

$$\begin{aligned}
dS_1(t) &= \mu_1 S_1(t)dt + S_1(t)\Big(\sigma_{11}dW_1(t) + \sigma_{12}dW_2(t)\Big)\\
dS_2(t) &= \mu_2 S_2(t)dt + S_2(t)\Big(\sigma_{21}dW_1(t) + \sigma_{22}dW_2(t)\Big)\,.
\end{aligned}$$

  The two price processes are correlated if $\sigma_{12} = \sigma_{21} \neq 0$. This setup can easily be extended to include n assets. Furthermore, the system of SDEs is linear.

- Linear SDE with stochastic volatility:
  The observed volatility for real existing price processes, such as stocks or bonds, is not constant as most models assume, but it is itself a stochastic process. In order to capture this observed fact, the price is modeled by two SDEs. The first SDE describes the logarithm of the price and the second SDE describes the evolution of the volatility $\sigma(t)$ over time. The following model is obtained:

$$\begin{aligned} dP(t) &= \mu dt + \sigma(t) dW_1(t) \\ P(0) &= P_0 \\ d\sigma(t) &= \kappa(\theta - \sigma(t))dt + \sigma_1 dW_2(t) \\ \sigma(0) &= \sigma_0 \,, \end{aligned}$$

  where θ is the average volatility, σ_1 a volatility, and κ the mean reversion rate of the volatility process $\sigma(t)$. If this model is used for stock prices, the transformation $P(t) = \ln(S(t))$ is useful. The two Brownian motions $dW_1(t)$ and $dW_2(t)$ are correlated, hence $\text{corr}[dW_1(t), dW_2(t)] = \rho$. This model captures the behavior of real existing prices better and its distribution of returns shows "fatter tails".

3.4.5 Nonlinear SDEs and Popular Nonlinear Pricing Models

In comparison with linear SDEs, nonlinear SDEs are less well understood. No general solution theory exists. And there are no explicit formulae for calculating the moments. In this section, we show some examples of nonlinear SDEs and their properties.

- Square Root processes:
  In general, a scalar square root process can be written as

$$dX(t) = f(t, X(t))dt + g(t, X(t))dW(t)$$

  with

$$\begin{aligned} f(t, X(t)) &= A(t)X(t) + a(t) \\ g(t, X(t)) &= B(t)\sqrt{X(t)} \,, \end{aligned}$$

  where $A(t)$, $a(t)$, and $B(t)$ are real scalars. The nonlinear mean reverting SDEs differ from the linear scalar equations by their nonlinear diffusion term.
  A mean reverting square root process is described by

$$\begin{aligned} dS(t) &= \kappa[\mu - S(t)]dt + \sigma\sqrt{S(t)}\, dW(t) \\ S(0) &= S_0 \,. \end{aligned}$$

  The process shows a less volatile behavior than its linear geometric counterpart and it has a non-central chi-square distribution. The process is

often used to model short-term interest rates or stochastic volatility processes for stock prices.
Another often used square root process is similar to the geometric Brownian motion, but with a square root diffusion term instead of the linear diffusion term. Its model is given by

$$dS(t) = \mu S(t)dt + \sigma\sqrt{S(t)}dW(t)\,, \quad S(0) = S_0\,.$$

Its mean and variance are:

$$\begin{aligned} \mathrm{E}[S(t)] &= S_0 e^{\mu t} \\ \mathrm{Var}[S(t)] &= \frac{\sigma^2 S_0}{\mu}\left(e^{2\mu t} - e^{\mu t}\right). \end{aligned}$$

In comparison with the geometric Brownian motion model, the variance of this model grows much more slowly over time.

- Mean reversion with log-normal distribution:
  Another widely used mean reversion model is obtained by

$$dS(t) = \kappa S(t)[\mu - \ln(S(t))]dt + \sigma S(t)dW(t)\,.$$

Using the transformation $P(t) = \ln(S(t))$ yields the linear mean reverting and normally distributed process $P(t)$:

$$dP(t) = \kappa\Big[\mu - \frac{\sigma^2}{2\kappa} - P(t)\Big]dt + \sigma dW(t)\,.$$

Because of the transformation, $S(t)$ is log-normally distributed. This model is used to model stock prices, stochastic volatilities, and electricity prices. Because $S(t)$ is log-normally distributed, $S(t)$ is always positive.

3.5 Partial Differential Equations and SDEs

In this section, we highlight the connection between stochastic differential equations and partial differential equations (PDEs).

Consider the stochastic process $X(.)$ with the deterministic initial value $x \in \mathbb{R}^n$ at the given initial time t,

$$X(t) = x\,,$$

which is governed by the stochastic differential equation

$$dX(s) = f(s, X(s))ds + g(s, X(s))dW(s) \quad \text{for} \quad s \geq t\,. \tag{3.29}$$

Here, $dW(s) \in \mathbb{R}^k$, and $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$, and $g : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^{n\times k}$.

We are interested in the stochastic "cost" functional which we associate with the stochastic process $X(.)$ over the time interval $[t, T]$

$$K(X(T)) + \int_t^T L(s, X(s))\, ds\,.$$

In particular, we want to compute its expected value which we denote by

$$F(t,x) = \mathrm{E}_{t,x}\Big[K(X(T)) + \int_t^T L(s, X(s))\, ds\Big]\,. \tag{3.30}$$

Note that the expectation operator $\mathrm{E}_{t,x}$ is the conditional expectation with respect to the given initial time t and the fixed initial state x.

Furthermore, we are also interested in the so-called "cost-to-go" function $F(t,x)$ for all starting times $t \leq T$ and all starting values $x \in \mathbb{R}^n$.

Example 3.12. We are interested in the average volatility of the stock market in a given time period $[0, T]$. The problem is to find the expected average volatility as described by

$$\begin{aligned}
&\mathrm{E}_{0,x}\Big[\int_0^T \frac{1}{T} x(t)\, dt\Big] \\
&dx(t) = \kappa(\theta - x(t))dt + \sigma\sqrt{x(t)}\, dW(t) \\
&x(0) = x_0\,,
\end{aligned}$$

where the SDE describes the dynamics of the market's volatility $x(t)$. One could for instance buy a derivative that gives a monetary payoff which equals the average volatility. Such instruments are known as volatility swaps.

The following result is of paramount importance: The cost-to-go function can be found by solving a partial differential equation!

Theorem 3.13. Feynman-Kač
The cost-to-go function $F(t,x)$ defined in (3.30) which is associated with the stochastic process $X(.)$ governed by the SDE (3.29) satisfies the PDE

$$\begin{aligned}
\frac{\partial F}{\partial t}(t,x) + \frac{\partial F}{\partial x}(t,x) f(t,x) + L(t,x) \\
+ \frac{1}{2}\mathrm{tr}\Big[g(t,x) g^T(t,x) \frac{\partial^2 F}{\partial x^2}(t,x)\Big] = 0
\end{aligned} \tag{3.31}$$

with the boundary condition

$$F(T,x) = K(x) \tag{3.32}$$

at the final time T.

The solution to the PDE problem is well-defined if the integrability conditions for the SDE are met.

This PDE and other similar PDEs will turn up in many applications where the problem is modeled by SDEs. This PDE is essential for finding solutions to derivative pricing or optimal control applications.

Proof of Theorem 3.13:
By definition, $F(t,x)$ is deterministic. Since integration and the expectation operator are linear, we can write

$$\mathrm{E}_{t,x}[F(T,X(T))] = F(t,x) + \mathrm{E}_{t,x}\Big[\int_t^T \frac{dF}{d\tau}(\tau,X(\tau))\,d\tau\Big]\,.$$

Reordering this equation, and using (3.30) and Itô's rule, we get

$$\begin{aligned}
&F(t,x)\\
&= \mathrm{E}_{t,x}\Big[K(X(T)) + \int_t^T L(\tau,X(\tau))\,d\tau\Big]\\
&= \mathrm{E}_{t,x}[F(T,X(T))]\\
&\quad - \mathrm{E}_{t,x}\Big[\int_t^T \Big\{\Big(\frac{\partial F}{\partial \tau} + \frac{\partial F}{\partial X}f + \frac{1}{2}\,\mathrm{tr}\Big\{gg^T\frac{\partial^2 F}{\partial X^2}\Big\}\Big)\,d\tau + \frac{\partial F}{\partial X}\,g\,dW(\tau)\Big\}\Big]\,,
\end{aligned}$$

where for ease of typesetting and reading, all of the identical arguments "$(\tau, X(\tau))$" have been suppressed.

First, we notice that for $t = T$ we get the claimed deterministic boundary condition

$$F(T,x) = K(x)\,.$$

Next, considering that the Brownian motion $W(.)$ is zero-mean and cancelling the boundary condition in the above equation yields the claimed partial differential equation in the first preliminary form

$$\mathrm{E}_{t,x}\Big[\int_t^T \Big\{\Big(L + \frac{\partial F}{\partial \tau} + \frac{\partial F}{\partial X}f + \frac{1}{2}\,\mathrm{tr}\Big\{gg^T\frac{\partial^2 F}{\partial X^2}\Big\}\Big)\,d\tau\Big\}\Big] = 0$$

and in the second preliminary form

$$\mathrm{E}_{t,x}\Big[L + \frac{\partial F}{\partial \tau} + \frac{\partial F}{\partial X}f + \frac{1}{2}\,\mathrm{tr}\Big\{gg^T\frac{\partial^2 F}{\partial X^2}\Big\}\Big] = 0 \quad \text{for all } \tau \in [t,T]$$

and in the final form

$$L(t,x) + \frac{\partial F}{\partial t}(t,x) + \frac{\partial F}{\partial x}(t,x)f(t,x) + \frac{1}{2}\,\mathrm{tr}\Big\{g(t,x)g^T(t,x)\frac{\partial^2 F}{\partial x^2}(t,x)\Big\} = 0\,.$$

This completes the proof of this theorem.

We can also investigate another closely related PDE problem:

$$\frac{\partial F}{\partial t}(t,x) + \frac{\partial F}{\partial x}(t,x) f^T(t,x) + L(t,x) + r(t,x)F(t,x) + \frac{1}{2}\mathrm{tr}\Big[g(t,x)g^T(t,x)\frac{\partial^2 F}{\partial x^2}(t,x)\Big] = 0$$
$$F(T,x) = K(X) \quad (3.33)$$

with an arbitrary scalar-valued function $r(.,.)$. This PDE often arises in problems for financial derivatives and is also attributed to Feynman and Kač. The related probabilistic solution to (3.33) is given by

$$\begin{aligned} F(s,x) &= \mathrm{E}_{t,x}\Big[e^{\int_t^T r(t,X(t))dt}K(T,X) + \int_t^T e^{\int_t^s r(t,X(t))dt}L(s,X)ds\Big] \\ dX(t) &= f(t,X(t))dt + g(t,X(t))dW \\ X(t) &= x\,. \end{aligned}$$

Example 3.14. Possessing all of the necessary tools to solve the problem of the value of a volatility swap, we now solve

$$\begin{aligned} F(0,x) &= \mathrm{E}_{0,x}\Big[\int_0^T \frac{1}{T}x(s)ds\Big] \\ dx(t) &= \kappa(\theta - x(t))dt + \sigma\sqrt{x(t)}dW \\ x(0) &= x_0\,. \end{aligned}$$

This leads to the PDE

$$F_t(t,x) + \kappa(\theta - x)F_x(t,x) + \frac{x}{T} + \frac{1}{2}\sigma^2 x F_{xx}(t,x) = 0$$
$$F(T,x) = 0\,.$$

We now have to solve the PDE and make an ansatz for $F(t,x)$ and compute the corresponding partial derivatives. This leads us to

$$\begin{aligned} F(t,x) &= a_1(t) + a_2(t)x(t) \\ F_t(t,x) &= \dot{a}_1(t) + \dot{a}_2(t)x(t) \\ F_x(t,x) &= a_2(t) \\ F_{xx}(t,x) &= 0\,, \end{aligned}$$

with the terminal conditions $a_1(T) = 0$ and $a_2(T) = 0$. Using the partial derivatives, we arrive at

$$\begin{aligned}
\dot{a}_1(t) + \dot{a}_2(t)x(t) + \kappa(\theta - x)a_2(t) + \frac{x(t)}{T} &= 0 \\
\dot{a}_1(t) + \kappa\theta a_2(t) + (\dot{a}_2(t) - \kappa a_2(t) + \frac{1}{T})x(t) &= 0 \\
\dot{a}_1(t) + \kappa\theta a_2(t) &= 0 \\
\dot{a}_2(t) - \kappa a_2(t) + \frac{1}{T} &= 0\,.
\end{aligned}$$

We have transformed the problem from finding a solution to a PDE into solving an ODE. The solution is

$$F(t,x) = \theta + \frac{\theta T}{\kappa}\Big(1 - e^{-\kappa T}\Big) + \frac{1}{\kappa T}\Big(1 - e^{-\kappa(t-T)}\Big)x(t)\,.$$

3.6 Solutions of Stochastic Differential Equations

In this part, we introduce three major methods to compute solution of SDEs. The first method is based on the Itô integral and has already been used for linear solutions. Secondly, we introduce numerical methods to compute path-wise solutions of SDEs. The third method is based on partial differential equations, where the problem of finding the probability density function of the solution is transformed into solving a partial differential equation.

3.6.1 Analytical Solutions of SDEs

Definition 3.15. *The stochastic process $X(t)$ governed by the stochastic differential equation*

$$\begin{aligned}
dX(t) &= f(t, X(t))dt + g(t, X(t))dW(t) \\
X(0) &= X_0
\end{aligned}$$

is explicitly described by the integral form

$$X(t,\omega) = X_0 + \int_0^t f(s, X(s))\,ds + \int_0^t g(s, X(s))\,dW(s)\,,$$

where the first integral is a path-wise Riemann integral and the second integral is an Itô integral.

In this definition, it is assumed that the functions $f(t, X(t))$ and $g(t, X(t))$ are sufficiently smooth in order to guarantee the existence of the solution $X(t)$.

There are several ways of finding analytical solutions. One way is to guess a solution and use the Itô calculus to verify that it is a solution for the SDE under consideration.

Example 3.16. We assume that the following nonlinear SDE

$$dX(t) = dt + 2\sqrt{X(t)}\, dW(t)\,,$$

has the solution

$$X(t) = (W(t) + \sqrt{X_0})^2\,.$$

We define $Z(t) = W(t)$ and thus $dZ(t) = dW(t)$ which is an SDE with $f = 0$ and $g = 1$. In order to verify this claim, we use the Itô calculus. We have $X(t) = \phi(Z)$ where $\phi(Z) = (W(t) + \sqrt{X_0})^2$, so that $\phi'(Z) = 2(Z(t) + \sqrt{X_0})$ and $\phi''(Z) = 2$. Using Itô's rule, we get

$$\begin{aligned} dX(t) &= \widetilde{f}(t, X)dt + \widetilde{g}(t, X)dW(t) \\ \widetilde{f}(t, X) &= \phi(W)'0 + \frac{1}{2}\phi''(W)1 = 1 \\ \widetilde{g}(T, X) &= \phi'(W)1 = 2(W(t) + \sqrt{X_0})\,. \end{aligned}$$

Since $X(t) = (W(t) + \sqrt{X_0})^2$ we know that $(W(t) + \sqrt{X_0}) = \sqrt{X(t)}$ and thus the Itô calculation generated the original SDE where we started at.

Another way of finding analytical solutions is to use the Itô calculus to transform the SDE in such a way that the resulting SDE can easily be integrated or that it already has a known analytic solution.

Example 3.17. As an example for this solution technique, we consider the following SDE

$$dX(t) = \Big(-\frac{1}{2}X(t) - \frac{1}{8X^3(t)}\Big)dt + \frac{1}{2X(t)}dW(t)\,.$$

For this rather unpleasant SDE, it is difficult to guess the solution. In order to simplify things, let us try the transformation $Y(t) = X^2(t)$, so that $\phi'(X(t)) = 2X(t)$ and $\phi''(X(t)) = 2$. Using Itô's rule, we get

$$dX(t) = \widetilde{f}(t, X(t))dt + \widetilde{g}(t, X(t))dW(t)$$

with

$$\begin{aligned} \widetilde{f}(t, X(t)) &= \phi(X(t))'\Big(-\frac{1}{2}X(t) - \frac{1}{8X^3(t)}\Big)dt + \frac{1}{2}\frac{1}{4X^2(t)}\phi''(X(t)) \\ &= -X^2(t) \\ \widetilde{g}(t, X(t)) &= \phi(X(t))'\frac{1}{2X(t)} = 1 \\ dY(t) &= -X^2(t)dt + dW(t) \\ &= -Y(t)dt + dW(t)\,. \end{aligned}$$

Hence, $Y(t)$ is the well-known Ornstein-Uhlenbeck process. With the transformation rule $X(t) = \sqrt{Y(t)}$ we have solved the original SDE.

3.6.2 Numerical Solution of SDEs

In some of the examples, we found explicit solutions for a given stochastic differential equation. However, most SDEs, especially nonlinear SDEs, do not have analytical solutions so that one has to resort to numerical approximation schemes in order to simulate sample paths of solutions to the given equation. We state here the two most common numerical procedures:

The simplest scheme is obtained by using a first-order approximation. This is called the Euler scheme

$$X(t_k) = X(t_{k-1}) + f(t_{k-1}, X(t_{k-1}))\Delta t + g(t_{k-1}, X(t_{k-1}))\Delta W(t_k)\,.$$

The Brownian motion term can be approximated as follows:

$$\Delta W(t_k) = \epsilon(t_k)\sqrt{\Delta t}\,,$$

where the $\epsilon(.)$ is a discrete-time Gaussian white process with mean 0 and standard deviation 1.

To get a somewhat better approximation, one can use the so-called Milstein scheme which includes second-order information

$$\begin{aligned} X(t_k) = X(t_{k-1}) &+ f(t_{k-1}, X(t_{k-1}))\Delta t + g(t_{k-1}, X(t_{k-1}))\Delta W(t_k) \\ &+ \frac{1}{2}\frac{\partial g}{\partial X}(t_{k-1}, X(t_{k-1}))[(\Delta W(t_k))^2 - \Delta t]. \end{aligned}$$

For the derivation and proof of the schemes the reader is referred to [26]. Both schemes converge to true sample paths if the step size in the numerical approximations are taken smaller and smaller. But higher-order schemes converge more quickly. This is important since, in practice, roundoff errors get in the way when the step size is taken too small. In order to compute the approximate solution of the SDE, we need to simulate a significant number of sample paths, since one sample path is just one realization of the unknown time-varying distribution.

3.6.3 Solutions of SDEs as Diffusion Processes

Introduction and Motivation

In the previous sections, some analytical and numerical examples of solutions for SDEs are given. However, these methods can only deliver one sample path of the given process, given a realization of a Brownian motion. Therefore, we are interested in knowing the statistical distribution of the possible outcomes of the stochastic process at all times. In other words, we want to know the transition probability $P(s, x, t, B)$ of the process.

We consider the stochastic differential equation

$$dX(t) = f(t, X(t))dt + g(t, X(t))dW(t)\,. \tag{3.34}$$

We know that the analytical solution of this general equation for $X(t)$ can only be found under certain conditions for $f(t, X(t))$ and $g(t, X(t))$. The question we want to answer here is the following: can we obtain the transition probability $P(s, x, t, B)$ of $X(t)$ directly from the stochastic differential equation (3.34), or rather from its drift and diffusion functions $f(t, X(t))$ and $g(t, X(t))$?

Transition Probability

We begin here by recalling the definition of the transition probability $P(s, x, t, B)$. This is the function we eventually are going to find and use to describe the solution of the stochastic differential equation (3.34).

Definition 3.18. *Let $X(t)$ be a Markov process. The function $P(s, x, t, B)$ is the conditional distribution of the probability $P(X(t) \in B \mid X(s) = x)$ and is called the transition probability. It is best described as the probability that the process is found inside the set B at time t, when at time $s < t$ it was found to be in state $X(s) = x$ (see Figure 3.1). The associated transition probability density $p(s, x, t, y)$ is given by*

$$P(s, x, t, B) = \int_B p(s, x, t, y)\, dy.$$

A Markov process is called shift-invariant if its transition probability $P(s, x, t, B)$ is stationary, i.e.,

$$P(s + u, x, t + u, B) \equiv P(s, x, t, B)\,.$$

In this case, the transition probability is only a function of x, $t - s$, and B and can be written in the following form

$$P(s, x, t, B) = P(t - s, x, B)\,,$$

such that $P(t, x, B)$ is the probability of transition of $X(t)$ into B over the time interval t, independently of where the time interval of length t lies.

Diffusion Processes

Generally speaking, a diffusion process is a kind of Markov process with continuous realizations whose transition probability $P(s, x, t, B)$ has certain properties for $t \downarrow s$. The simplest example of a diffusion process is the Brownian motion.

Definition 3.19. *A Markov process X with values in R^d is called a diffusion process, if its transition probability $P(s, x, t, B)$ satisfies the following conditions for $s \in [t_o, T)$, $X \in R^d$, and $\varepsilon > 0$*

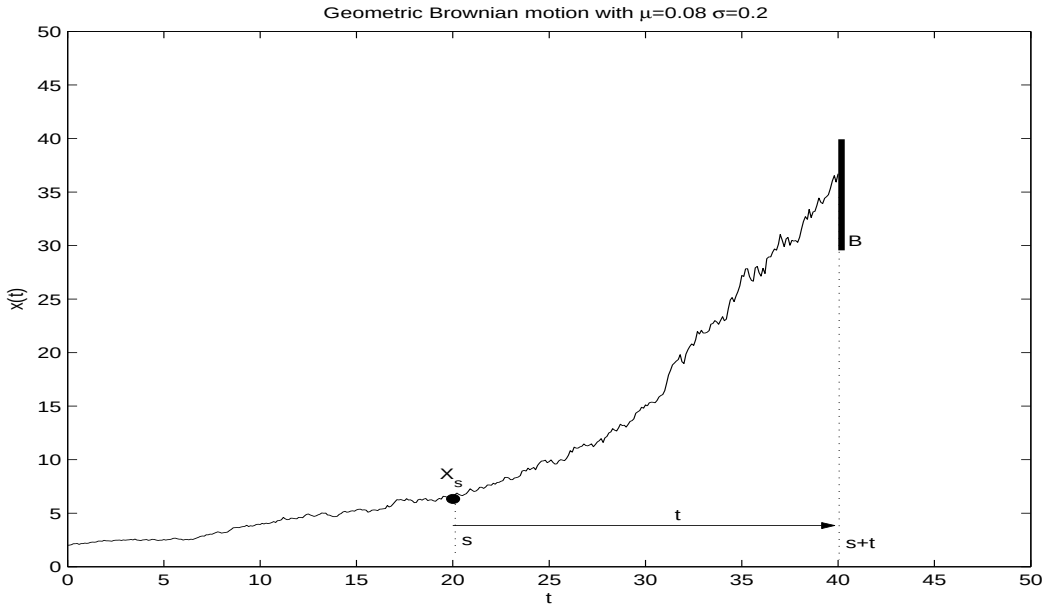

Fig. 3.1. Transition probability $P(s, X, t, B)$.

a) Continuity:

$$\lim_{t \downarrow s} \frac{1}{t-s} \int_{|y-x|>\varepsilon} P(s,x,t,dy) = 0\,.$$

b) There exists a function $f(t,x)$ in R^d with

$$\lim_{t \downarrow s} \frac{1}{t-s} \int_{|y-x|\leq\varepsilon} (y-x) P(s,x,t,dy) = f(s,x)\,.$$

c) There exists a symmetric $d \times d$ matrix $\Sigma(t,x)$ with

$$\lim_{t \downarrow s} \frac{1}{t-s} \int_{|y-x|\leq\varepsilon} (y-x)(y-x)^T P(s,x,t,dy) = \Sigma(s,x)\,.$$

The functions $f(t,x)$ and $\Sigma(t,x) = g(t,x)g^T(t,x)$ are called the coefficients of the diffusion process, f is the drift vector and Σ the diffusion matrix.

Conditions a), b), and c) have the following meanings (assuming the first and second moments of X exist, the index denoting the time at which the process is evaluated, and at time s, we have the boundary condition $X_s = x$):

Condition a) makes it improbable for the process X to have large changes in values in a short time interval. It thereby rules out any Poisson type effects in the diffusion process:

$$P(\mid X_t - X_s \mid \leq \varepsilon \mid X_s = x) = 1 - o(t-s).$$

For condition b) we can write

$$\lim_{t \downarrow s} E_{s,x}(X_t - X_s) = f(s,x)(t-s) + o(t-s)$$

and with condition c)

$$\lim_{t \downarrow s} E_{s,x}(X_t - X_s)(X_t - X_s)^T = \Sigma(s,x)(t-s) + o(t-s)$$

or

$$\lim_{t \downarrow s} \mathrm{Cov}_{s,x}(X_t - X_s) = \Sigma(s,x)(t-s) + o(t-s)\,.$$

We can now state that f is the average speed of the stochastic movement and Σ describes the magnitude of the fluctuation $X_t - X_s$ around the mean value of X with respect to the transition probability $P(s,x,t,B)$. We now choose an arbitrary matrix $g(t,x)$ such that $\Sigma = gg^T$ and can write the following equation:

$$X_t - X_s = f(s,X_s)(t-s) + g(s,X_s)\xi,$$

where $E_{s,x}\xi = 0$ and $\mathrm{Cov}_{s,x}\xi = (t-s)I$. We can replace this by a Brownian motion, as we know that $W_t - W_s$ follows the distribution $\mathcal{N}(0,(t-s)I)$, i.e., exactly the distribution of ξ that we were looking for. So,

$$X_t - X_s = f(s,X_s)(t-s) + g(s,X_s)(W_t - W_s)$$

or, in differential notation,

$$dX_t = f(t,X_t)dt + g(t,X_t)dW_t\,. \tag{3.35}$$

This closely resembles equation (3.34) of the problem description. Note that this form of the SDE (3.35) has been derived directly from the transition probabilities and the characteristics of diffusion processes without knowing the general solution of (3.34) in advance. This goes to show how the same process can be described either through transition probabilities or through a stochastic differential equation.

Fokker-Planck Equation

After having seen the link between the transition probability and the stochastic process above, it is now necessary to define a method of directly stating the transition probability as a function of the process parameter functions f and B. An analytical approach is the following:

Let X be the solution of a d-dimensional diffusion process

$$dX(t) = f(t,X(t))dt + g(t,X(t))dW(t)\,,$$

whose drift term f and diffusion coefficient g are sufficiently smooth functions. Then the transition probabilities $P(s,x,t,B) = P(X(t) \in B \mid X(s) = x)$ are given as the solution of the equations

$$\begin{aligned} P_s(s,x,t,B) &+ f^T(s,x)P_x(s,x,t,B) \\ &+ \frac{1}{2}\mathrm{tr}\{g(s,x)g^T(s,x)P_{xx}(s,x,t,B)\} = 0 \\ P(t,x,t,B) &= I_B(X(t))\,, \end{aligned} \tag{3.36}$$

where I_B is the indicator function of the set B. We further know that

$$P(s,x,t,B) = E_{s,x}I_B(X(t)) = P(X(t) \in B \mid X(s) = x)\,.$$

The transition probability $P(s,x,t,B)$ has the density $p(s,x,t,y)$ and we can directly write the so-called *Kolmogorov backward equation* for the transition probability density function $p(s,x,t,y)$

$$\begin{aligned} p_s(s,x,t,y) &+ f^T(s,x)p_x(s,x,t,y) \\ &+ \frac{1}{2}\mathrm{tr}\{g(s,x)g^T(s,x)p_{xx}(s,x,t,y)\} = 0 \\ \lim_{s\uparrow t} p(s,x,t,y) &= \delta(y-x)\,. \end{aligned} \tag{3.37}$$

The reason for calling this equation a backward equation is that the differential operators are considered with regard to the backward variables (s,x); in other words, the differential equation gives the backward evolution with respect to the initial state, given the final state (t,y).

Finally, the *Fokker-Planck equation* gives the forward evolution with respect to the final state (t,y) starting at the initial state (s,x) and therefore, it is called a *forward equation*.

$$\begin{aligned} p_t(s,x,t,y) &+ f^T(t,y)p_y(s,x,t,y) \\ &- \frac{1}{2}\mathrm{tr}\{g(t,y)g^T(t,y)p_{yy}(s,x,t,y)\} = 0 \\ \lim_{t\downarrow s} p(s,x,t,y) &= \delta(y-x). \end{aligned}$$

Example 3.20. For the Wiener process, the forward equation for the homogeneous transition density is

$$\begin{aligned} p(t,x,y) &= (2\pi t)^{-d/2}e^{-|y-x|^2/2t} \\ \frac{\partial p}{\partial t} &= \frac{1}{2}\sum_{i=1}^{d}\frac{\partial^2 p}{\partial y_i^2}\,. \end{aligned}$$

Example 3.21. The scalar linear differential equation

$$dX_t = (AX(t) + a)dt + bdW(t), \quad X(0) = c, \quad t \geq 0,$$

has the following solution

$$X(t) = ce^{At} + \frac{a}{A}(e^{At} - 1) + b\int_0^t e^{A(t-s)}dW(s)\,.$$

This includes the Ornstein-Uhlenbeck process ($a = 0$), a deterministic differential equation ($b = 0$), and the standard Brownian motion ($A = a = 0$, $b = 1$, $c = 0$) as special cases. With $b \neq 0$ we can find a solution for the transition probability density $p(s, x, t, y)$ through the Fokker-Planck equation

$$\frac{\partial}{\partial t}p(s,x,t,y) + (Ay + a)\frac{\partial}{\partial y}p(s,x,t,y) - \frac{1}{2}b^2\frac{\partial^2}{\partial y^2}p(s,x,t,y) = 0\,.$$

As boundary conditions we assume that p vanishes (including all its partial derivatives) for $|\,x\,|\to\infty$ and $|\,y\,|\to\infty$. We know that

$$p(s,x,t,B) = (2\pi K_t(s,x))^{-1/2}\exp\left(-\frac{(y - m_t(s,x))^2}{2K_t(s,x)}\right)$$

holds with

$$\begin{aligned} m_t(s,x) &= xe^{A(t-s)} + \frac{a}{A}(e^{A(t-s)} - 1) \\ K_t(s,x) &= \frac{b^2}{2A}(e^{2A(t-s)} - 1)\,, \end{aligned}$$

i.e., $p(s, x, t, B)$ is the density of the normal distribution $\mathcal{N}(m_t(s, x), K_t(s, x))$.

3.7 Stability

3.7.1 Introduction

Every autonomous linear time-invariant dynamic system

$$\dot{x}(t) = Ax(t)$$

has either exactly one equilibrium state or it has infinitely many equilibrium states. If all of the eigenvalues of the matrix A have strictly negative real parts, $x(t) \equiv 0$ is the only equilibrium state. This equilibrium state is asymptotically stable because the transition matrix asymptotically vanishes:

$$\lim_{t\to\infty} e^{At} = 0\,.$$

In other words, the state vector $x(t)$ asymptotically vanishes for every initial state $x(0) = x_0 \in R^n$. In this case, we say, the linear time-invariant system is asymptotically stable.

If at least one of the eigenvalues of the matrix A is zero, then every vector x in the null space of A, $x \in N(A)$, is an equilibrium state. The system may be unstable or stable (but not asymptotically stable) depending on the real part of the dominant pole and the multiplicity of the vanishing eigenvalue of A.

The stability properties of a linear time-varying dynamic system is a more complicated issue because the (quasi-stationary) stability properties can change as a function of the time t.

An autonomous nonlinear time-invariant dynamic system

$$\dot{x}(t) = f(x(t)) \tag{3.38}$$

may have an arbitrary number of equilibrium states. Each one of them is characterized by $f(x) = 0$. Some of these equilibrium states may be asymptotically stable, or unstable, or marginally stable. Also, the system may have no equilibrium state at all.

Example 3.22. Think of a roller-coaster where the motion of the cart is subject to friction. The state vector consists of the position of the cart and the cart's velocity. Every bottom of a strictly convex (local) valley corresponds to an asymptotically stable equilibrium; every top of a strictly concave (local) hill corresponds to an unstable equilibrium; and every point on a level plateau corresponds to a marginally stable equilibrium.

The underlying definition of stability of an equilibrium state is as follows:

Definition 3.23. *An equilibrium state x^* of the nonlinear time-invariant system (3.38) is called asymptotically stable if for some value $r > 0$ and for all initial states $x(0)$ with $\|x(0) - x^*\| < r$, we get asymptotic reverting, i.e., $\lim_{t\to\infty} x(t) = x^*$; it is called a stable equilibrium if for some value $R > 0$, the state remains bounded by $\|x(t) - x^*\| < R$ for all $t > 0$; and it is called an unstable equilibrium otherwise.*

Of course, again, for a time-varying nonlinear dynamic system, the situation is more complicated. This is not pursued here.

For a stochastic autonomous nonlinear time-invariant dynamic system

$$dX(t) = f(X(t))dt + g(X(t))dW(t)\,,$$

the situation is even more complex due to the omnipresent white noise driving the system: if the state x^* is an asymptotically stable equilibrium (of the deterministic system), there is no guarantee that the state will remain close to x^* because the influences of the noise may eventually drive the system "over the nearest hill".

Therefore, we need to refine the concept of stability and we need to define different stability measures such as asymptotic stability, exponential stability, and stability in probability.

3.7.2 Moment Method for Stochastic Systems

The moment method for stochastic systems is a way to study the stability of a stochastic system whose solution is already known. As we shall see, this might be a disadvantage of the method, but it provides a very intuitive view on stability. It is an obvious expansion of the known deterministic stability measure and it is very interesting nontheless.

Consider the stochastic differential equation

$$dX(t) = f(t, X(t))dt + g(t, X(t))dW(t), \quad X(t_0) = x_0. \tag{3.39}$$

In the following, we assume $f(t, 0) \equiv g(t, 0) \equiv 0$ such that the trivial solution $x(t) = 0$ holds for $x_0 = 0$.

Several different stability statements are possible:

Definition 3.24. *The process is called*

1. ***p-stable*** *if, for each $\epsilon > 0$, there is a $\delta > 0$ such that*

$$\sup_{t_0 \le t < \infty} E|X(t)|^p \le \epsilon\,, \quad \textit{for all } X(t_0) = x_0\,,\ |x_0| \le \delta.$$

   This stability description for the p-th moment of a stochastic system comes close to the basic stability description given in Definition 3.23. Here, we want to know about stability of the p-th moment of a stochastic system, i.e., for $p = 1$ we regard the special case of stability of the expected value of the process.

2. ***asymptotically p-stable*** *if it is p-stable and there is an $\epsilon > 0$ such that we have*

$$\lim_{t \to \infty} E|X(t)|^p = 0\,, \quad X(t_0) = x_0\,.$$

   This is a stronger type of stability, as in this case, when the system is disturbed from its nominal position, we know that it will converge back to this point as $t \to \infty$. Again, for stochastic systems we can look at the asymptotic stability of any p-th moment with this method.

3. ***exponentially p-stable*** *if there are two positive constants, c_1, c_2, such that*

$$E|X(t)|^p \le c_1 |x_0|^p e^{-c_2(t-t_0)}\,, \quad X(t_0) = x_0\,.$$

   Expressed in words, exponentially stable means that the state vector of an exponentially stable system converges to the origin faster than an exponential function. We call c_2 the rate of exponential convergence.

4. ***stable in probability*** *if for arbitrary $\epsilon_1 > 0$, $\epsilon_2 > 0$ there is a $\delta > 0$ such that*

$$P\Big\{ \sup_{t_0 \le t < \infty} |X(t)| \le \epsilon_1 \Big\} \ge 1 - \epsilon_2\,, \quad \textit{for all } X(t_0) = x_0\,,\ |x_0| \le \delta\,.$$

   Here, we know that the state vector will stay within certain limits for all times $t \to \infty$ with a given probability.

We call he special cases $p = 1$ and $p = 2$ mean-stability and mean-square stability, respectively.

The difficulty with this definition of stability is that we need to know the solution of equation (3.39), before being able to investigate stability. A second method, found 1892 by A. M. Lyapunov, allows to prove stability without necessarily knowing the solution of the given system.

3.7.3 Lyapunov's Second Method

Lyapunov Method for Deterministic Systems

In the first step, we want to discuss the stability of deterministic systems. We use the stability theory of Lyapunov which can then be expanded to suit the special needs of stochastic systems.

We begin with a deterministic differential equation for the d-dimensional state $x(t)$

$$\dot{x}(t) = f(t, x(t)), \quad x(t_0) = x_0, \quad t \geq t_0\,. \tag{3.40}$$

We set $f(t, 0) = 0$ such that the solution $x(t) \equiv 0$ holds for $x_0 = 0$. Thus, $x = 0$ is an equilibrium state. We call this equilibrium stable if for every $\epsilon > 0$ there exists a $\delta(\epsilon, t_0) > 0$ such that $\sup_{t_0 \leq t < \infty} |x(t)| \leq \epsilon$ as long as $x(t_0) = x_0$ and $|x_0| \leq \delta$.

Definition 3.25. *We call a scalar function $v(x)$ positive-definite if $v(0) = 0$ and $v(x) > 0$ for all $x \neq 0$. The function is defined in the spherical region $U_h = \{x : |x| \leq h\} \subset R^d$, $h > 0$. We call the function radially unbounded, if*

$$\inf_{t \geq t_0} v(t, x) \to \infty \quad for\ |x| \to \infty\,.$$

We assume that $x(t)$ is a solution to equation (3.40) and $v(t, x)$ is a positive-definite function with continuous partial first-order derivatives. We further regard $V(t) = v(t, x(t))$ with its derivative

$$\frac{dV(t)}{dt} = \frac{\partial v}{\partial t} + \sum_{i=1}^{d} \frac{\partial v}{\partial x_i} f_i(t, x(t)).$$

If $\frac{dV(t)}{dt} \leq 0$, we know that $x(t)$ follows a path such that the value of $V(t)$ does not increase, i.e., the "distance" of $x(t)$ to its equilibrium "measured" by $v(t, x(t))$ does not increase.

More generally, we can say that if there exists a positive-definite function $v(t, x)$ such that the first-order derivative satisfies

$$\frac{dv(t, x)}{dt} = \frac{\partial v}{\partial t} + \sum_{i=1}^{d} \frac{\partial v}{\partial x_i} f_i(t, x) \leq 0$$

on the trajectories the of differential equation (3.40), i.e., $v(t, x(t))$, the system's equilibrium is called stable. A function $v(t, x)$ which fulfills the conditions for a stability proof is called a Lyapunov function.

Lyapunov Method for Stochastic Systems

We now need to expand the stability analysis according to Lyapunov to the world of stochastic systems. However, the first problem is to create a suitable definition of stability for stochastic systems and to change the definition of the Lyapunov function $v(t,x)$ and the stability condition $\frac{dv}{dt} \leq 0$ accordingly.

We consider the stochastic differential equation

$$dX(t) = f(t,X(t))dt + g(t,X(t))dW(t), \quad X(t_0) = c, \quad t \geq t_0, \tag{3.41}$$

whose solution $X(t)$ is a d-dimensional diffusion process with drift $f(t,x)$ and diffusion matrix $\Sigma(t,x) = g(t,x)g^T(t,x)$. We set $f(t,0) = 0$, $g(t,0) = 0$ for all $t \geq t_0$, such that the equilibrium $X(t) \equiv 0$ is the solution of the differential equation with $c = 0$.

Let $X(t)$ be the solution to equation (3.41) and $v(t,x)$ a positive-definite function with continuous partial differentials. The process $V(t) = v(t,X(t))$ then has a stochastic differential according to the Itô differential operator

$$L(\cdot) = \frac{\partial(\cdot)}{\partial t} + \frac{\partial(\cdot)}{\partial x} f(t,x) + \frac{1}{2}\mathrm{tr}\Big\{ g(t,x)g^T(t,x)\frac{\partial^2(\cdot)}{\partial x^2} \Big\}.$$

The differentiation of $V(t,X_t)$ yields

$$dV(t) = (Lv(t,X(t)))dt + \sum_{i=1}^{d}\sum_{j=1}^{m} v_{x_i}(t,X(t))g_{ij}(t,X(t))dW(t)^j\,.$$

For a stable equilibrium, $V(t)$ does not increase, i.e., $dV(t) \leq 0$. Due to the last term with the stochastic property, this cannot be easily guaranteed in most cases. Therefore, we will only demand that $X(t)$ stays near the equilibrium on average. This results in

$$E(dV(t)) \leq 0\,.$$

With $E(dV(t)) = E(Lv(t,X(t))dt)$, we finally get the condition for stochastic stability in the Lyapunov sense:

$$Lv(t,x) \leq 0\,, \quad \text{for all} \quad t \geq t_0\,.$$

We call the function $v(t,x)$ the Lyapunov function belonging to a particular equilibrium state of the stochastic differential equation (3.41).

Notes and Comments

The literature on stochastic calculus and stochastic differential equations is vast. Fairly straightforward introductions can be found in [6], [23], and [27]. Mathematically more rigorous textbooks are [28], [26], and [3].

4

Model-Based Filtering

She: How often does a Kalman filter need to be cleaned?
He: Never.
She: Why not?
He: Because it is optimal.
She: Can you get me one for our coffee machine?
He: No. It is something theoretical.
Dialogue of an unknown couple

Filtering and estimation are two closely related concepts. When a stochastic dynamic system is modeled, we usually are not able to measure every state variable of the system. In the deterministic case, the Luenberger observer is used for reconstructing the missing information from the measurements and the system dynamics. In almost every real-world application, noise is present in the system itself as well as in the measurements. If we need an estimation of the state of such a noisy system, we make use of the concepts of filtering. We therefore refer to a filter as a state estimator for a stochastic dynamic system. Naturally, we demand some favorable properties for the filter. Most importantly, we want the estimation to bias-free, i.e., the estimation error should have zero expectation in order to prevent any systematic errors. In addition, the scattering of the estimate around the true value of the state should be as small as possible in order to have more confidence in the estimations. This is achieved by minimizing the variance of the estimation error.

Filtering techniques can also be used for estimating unknown parameters of a deterministic or stochastic dynamic system.

In this chapter, we shall use stochastic processes with independent increments $w_1(.)$ and $w_2(.)$ at the input and the output, respectively, of a dynamic system. We shall switch back and forth between the mathematically precise description of these (normalized) Brownian motions by their increments and the sloppy description by the corresponding white noises $\dot{v}(.)$ and $\dot{r}(.)$, respectively, which is preferred in engineering circles

$$Q^{1/2}(t)dw_1(t) = [\dot{v}(t) - \overline{u}(t)]dt \tag{4.1}$$

$$R^{1/2}(t)dw_2(t) = [\dot{r}(t) - \overline{r}(t)]dt\,, \tag{4.2}$$

where $Q^{1/2}(t)$ and $R^{1/2}(t)$ are the volatility parameters.

Furthermore, we abandon the rather cumbersome notation of stochastic processes by upper-case letters.

4.1 Linear Filtering

By linear filtering, we refer to problems with linear system dynamics. Consequently, the dynamics of the filter are linear as well. The concept of model-based filtering was published by Kalman and Bucy in the early 1960s.

4.1.1 The Kalman Filter

Continuous-time version of the Kalman filter

Consider the following linear time-varying dynamic system of order n which is driven by the m-vector-valued white noise $\dot{v}(.)$. Its initial state $x(t_0)$ is a random vector ξ and its p-vector-valued output $y(.)$ is corrupted by the additive white noise $\dot{r}(.)$:

System description in the mathematically precise form:

$$\begin{aligned} dx(t) &= A(t)x(t)dt + B(t)dv(t) \\ &= [A(t)x(t) + B(t)\overline{u}(t)]dt + B(t)Q^{1/2}(t)dw_1(t) && (4.3) \\ x(t_0) &= \xi && (4.4) \\ y(t)dt &= C(t)x(t)dt + dr(t) \\ &= [C(t)x(t) + \overline{r}(t)]dt + R^{1/2}(t)dw_2(t)\,, && (4.5) \end{aligned}$$

where

$$\begin{aligned} \xi &: \mathcal{N}(x_0, \Sigma_0) && (4.6) \\ dv(t) &: \mathcal{N}(\overline{u}(t)dt, Q(t)dt) && (4.7) \\ dr(t) &: \mathcal{N}(\overline{r}(t)dt, R(t)dt)\,. && (4.8) \end{aligned}$$

System description in the engineer's form[1]:

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)\dot{v}(t) && (4.9) \\ x(t_0) &= \xi && (4.10) \\ y(t) &= C(t)x(t) + \dot{r}(t)\,, && (4.11) \end{aligned}$$

where

$$\begin{aligned} \xi &: \mathcal{N}(x_0, \Sigma_0) && (4.12) \\ \dot{v}(t) &: \mathcal{N}(\overline{u}(t), Q(t)) && (4.13) \\ \dot{r}(t) &: \mathcal{N}(\overline{r}(t), R(t))\,. && (4.14) \end{aligned}$$

Note that in the last two lines, we have deleted the factor $\frac{1}{dt}$ which ought to multiply $Q(t)$ and $R(t)$. For white noises, it is customary to only notate the "covariance" or "intensity" parameters. For stationary white noises, Q and R are simply the spectral density matrices.

[1] Of course, an engineer would not use symbols such as $\dot{v}$ and $\dot{r}$ to denote white noise processes. — Just consider the derivate dots as concessions to mathematicians!

The random initial state ξ and the white noise processes $\dot{v}$ and $\dot{r}$ (or the normalized Wiener processes w_1 and w_2, respectively) are assumed to be mutually independent. In addition, we impose the following requirements:

$$\Sigma_0 = \Sigma_0^T \geq 0 \tag{4.15}$$
$$Q(t) = Q(t)^T \geq 0 \tag{4.16}$$
$$R(t) = R(t)^T > 0\,. \tag{4.17}$$

These requirements are pretty obvious. Requiring that the matrix $R(t)$ must be positive-definite means that the measurement $y(t)$ cannot contain any perfectly uncorrupted information.

The filtering problem is stated as follows: Find the optimal filter with the state vector $\widehat{x}$ which is optimal in the following sense:

- The state estimation is bias-free:
  $E\{x(t) - \widehat{x}(t)\} \equiv 0$.
- The error of the state estimate has infimal covariance matrix:
  $\Sigma_{opt}(t) \leq \Sigma_{subopt}(t)$.

Theorem 4.1. The Kalman Filter
With the above-mentioned assumptions that ξ, v, and r are mutually independent and Gaussian, the optimal filter is the following linear dynamic system:

$$\begin{aligned}\dot{\widehat{x}}(t) &= A(t)\widehat{x}(t) + B(t)\overline{u}(t) + H(t)[y(t) - \overline{r}(t) - C(t)\widehat{x}(t)] \\ &= [A(t) - H(t)C(t)]\widehat{x}(t) + B(t)\overline{u}(t) + H(t)[y(t) - \overline{r}(t)] \qquad (4.18)\\ \widehat{x}(t_0) &= x_0 \qquad (4.19)\end{aligned}$$

with

$$H(t) = \Sigma(t)C^T(t)R^{-1}(t)\ , \tag{4.20}$$

where $\Sigma(t)$ is the covariance matrix of the error $\widehat{x}(t)$–$x(t)$ of the state estimate $\widehat{x}(t)$ satisfying the following matrix Riccati differential equation:

$$\begin{aligned}\dot{\Sigma}(t) &= A(t)\Sigma(t) + \Sigma(t)A^T(t) \\ &\quad - \Sigma(t)C^T(t)R^{-1}(t)C(t)\Sigma(t) + B(t)Q(t)B^T(t) \qquad (4.21)\\ \Sigma(t_0) &= \Sigma_0\ . \qquad (4.22)\end{aligned}$$

Remark 4.2. If ξ, v, or r are not Gaussian, the Kalman filter is the best *linear* filter, but there may be a better *nonlinear* one.

A simple proof of Theorem 4.1:
The estimation error $e(t) = x(t) - \widehat{x}(t)$ satisfies the differential equation

$$\begin{aligned}\dot{e} &= \dot{x} - \dot{\widehat{x}} \\ &= Ax + B\overline{u} + B(\dot{v}-\overline{u}) - [A-HC]\widehat{x} - B\overline{u} - HCx - H(\dot{r}-\overline{r}) \\ &= [A-HC]e + B(\dot{v}-\overline{u}) - H(\dot{r}-\overline{r}) \ . \end{aligned} \tag{4.23}$$

Since the white noise processes $\dot{v}$ and $\dot{r}$ are uncorrelated, the covariance matrix $\Sigma(t)$ of the state estimation errer $e(t)$ is governed by the matrix differential equation

$$\begin{aligned}\dot{\Sigma} &= [A-HC]\Sigma + \Sigma[A-HC]^T + BQB^T + HRH^T \qquad (4.24)\\ &= A\Sigma + \Sigma A^T + BQB^T - \Sigma C^T R^{-1} C\Sigma \\ &\quad + [H-\Sigma C^T R^{-1}]\, R\, [H-\Sigma C^T R^{-1}]^T \\ &\geq A\Sigma + \Sigma A^T + BQB^T - \Sigma C^T R^{-1} C\Sigma \ . \qquad (4.25)\end{aligned}$$

Obviously, $\dot{\Sigma}(t)$ is infimized at all times t for

$$H(t) = \Sigma(t)C^T(t)R^{-1}(t) \ . \tag{4.26}$$

Hence, its integral $\Sigma(t)$ is infimized at all times as well. The optimal error covariance matrix of the state estimation errer $e(t)$ thus satisfies the following matrix Riccati differential equation:

$$\begin{aligned}\dot{\Sigma}(t) &= A(t)\Sigma(t) + \Sigma(t)A(t)^T + B(t)Q(t)B(t)^T \\ &\quad - \Sigma(t)C(t)^T R(t)^{-1} C(t)\Sigma(t) \qquad (4.27)\\ \Sigma(t_0) &= \Sigma_0 \ . \qquad (4.28)\end{aligned}$$

A more sophisticated proof of Theorem 4.1:
Consider again the differential equation (4.24) for the covariance matrix $\Sigma(t)$ of the state estimation error $e(t)$:

$$\dot{\Sigma} = [A-HC]\Sigma + \Sigma[A-HC]^T + BQB^T + HRH^T \ .$$

Its first differential with respect to H, at some value of H, with increment dH can be formulated as follows:

$$\begin{aligned}d\dot{\Sigma}(H, dH) &= \frac{\partial \dot{\Sigma}(H)}{\partial H} dH \\ &= -dHC\Sigma - \Sigma C^T dH^T + dHRH^T + HRdH^T \\ &= U[HR-\Sigma C^T]T dH \ , \end{aligned} \tag{4.29}$$

where T is the matrix transposing operator and U is the operator which adds the transposed matrix to its matrix argument[2]. In order for $\dot{\Sigma}(H)$ to

[2] $T : N \mapsto N^T$ and $U : M \mapsto M + M^T$, respectively.

have an infimum at some value of the matrix H, it is necessary that the first derivative vanishes at H [4]. This is satisfied if and only if the matrix in the square brackets vanishes, i.e., for

$$H(t) = \Sigma(t)C^T(t)R^{-1}(t) \ ,$$

where, again, $\Sigma(t)$ is the solution of the Riccati differential equation (4.27) with the boundary condition (4.28).

If the dynamic system $[A, B, C]$ is time-invariant, and if the two intensity matrices Q and R are constant, and if the system $[A, C]$ is completely observable [16, Ch. 4.7], the solution of the matrix Riccati differential equation becomes asymptotically constant (even if A is not a stability matrix): $\lim_{t\to\infty} \Sigma(t) = \Sigma_\infty$. For Kalman filters operating over very long time intervals, it is therefore attractive to use the time-invariant Kalman filter:

Corollary 4.3. The Time-Invariant Kalman Filter
For $t \in [0, \infty)$, in the time-invariant/quasi-stationary case, use the following time-invariant Kalman filter:

$$\begin{aligned} \dot{\widehat{x}}(t) &= A\widehat{x}(t) + B\overline{u}(t) + H[y(t) - \overline{r}(t) - C\widehat{x}(t)] \\ &= [A - HC]\widehat{x}(t) + B\overline{u}(t) + H[y(t) - \overline{r}(t)] \qquad (4.30) \\ \widehat{x}(t_0) &= x_0 \qquad (4.31) \end{aligned}$$

with

$$H = \Sigma_\infty C^T R^{-1} \ , \qquad (4.32)$$

where $\Sigma_\infty \geq 0$ is the covariance matrix of the error $\widehat{x}(t) - x(t)$ of the state estimate $\widehat{x}(t)$ satisfying the following algebraic matrix Riccati equation:

$$0 = A\Sigma_\infty + \Sigma_\infty A^T - \Sigma_\infty C^T R^{-1} C\Sigma_\infty + BQB^T \ . \qquad (4.33)$$

Note that the covariance matrix Σ_∞ of the state estimation error is positive-definite (rather than only positive-semidefinite), if and only if the system $[A, B]$ is completely controllable [16, Ch. 4.6].

Example 4.4. System of first order
Consider the following stochastic system of first order with the random initial condition $\xi : \mathcal{N}(x_0, \Sigma_0)$ and the uncorrelated white noises $\dot{v} : \mathcal{N}(0, Q)$ and $\dot{r} : \mathcal{N}(0, R)$:

$$\begin{aligned} \dot{x}(t) &= ax(t) + b[u(t) + \dot{v}(t)] \\ x(0) &= \xi \\ y(t) &= x(t) + \dot{r}(t) \, . \end{aligned}$$

Find the time-invariant Kalman filter!

The resulting time-invariant Kalman filter is described by the following equations:

$$\dot{\widehat{x}}(t) = -\sqrt{a^2 + b^2 \frac{Q}{R}}\, \widehat{x}(t) + bu(t) + \left(a + \sqrt{a^2 + b^2 \frac{Q}{R}}\right) y(t)$$
$$\widehat{x}(0) = x_0 \ .$$

Notice that the pole of the Kalman filter is at the left of $-|a|$ in the complex plane, irrespective of whether the given stochastic system is asymptotically stable ($a < 0$) or unstable ($a > 0$). In the former case, the filter gain H is smaller than in the latter. Also note that the ratio $\frac{Q}{R}$ is relevant rather than the absolute sizes of the intensities Q and R.

Continuous-time/discrete-time version of the Kalman filter

In the continuous-time/discrete-time problem we consider the case of continuous-time system dynamics and discrete-time measurements. Therefore, we need a discrete-time version of the Kalman filter for digital signal processing. The continuous-time measurement (4.11) which is corrupted by the white noise error $\dot{r}(t)$ with infinite covariance,

$$y(t) = C(t)x(t) + \dot{r}(t) \ ,$$

must be replaced by an "averaged" discrete-time measurement

$$y_k = C_k x(t_k) + \widetilde{r}_k \tag{4.34}$$
$$\widetilde{r}_k \ : \ \mathcal{N}(\overline{r}_k, R_k) \tag{4.35}$$
$$\mathrm{Cov}(\widetilde{r}_i, \widetilde{r}_j) = R_i \delta_{ij} \ . \tag{4.36}$$

At any sampling time t_k we consider the state estimate $\widehat{x}(t_k|t_{k-1})$ and the corresponding error covariance matrix $\Sigma(t_k|t_{k-1})$ before the new measurement y_k has been processed and the state estimate $\widehat{x}(t_k|t_k)$ and the corresponding error covariance matrix $\Sigma(t_k|t_k)$ after the new measurement y_k has been processed. — As a consequence, the continuous-time/discrete-time Kalman filter alternatingly performs update steps at the sampling times t_k processing the latest measurement information y_k and open-loop extrapolation steps between the times t_k and t_{K+1} with no measurement information available:

Lemma 4.5. The Continuous-Time/Discrete-Time Kalman Filter

Update at time t_k:

$$\begin{aligned}\widehat{x}(t_k|t_k) = \widehat{x}(t_k|t_{k-1}) & \\ + \Sigma(t_k|t_{k-1}) C_k^T \left[C_k \Sigma(t_k|t_{k-1}) C_k^T + R_k\right]^{-1} & \\ \times \left[y_k - \overline{r}_k - C_k \widehat{x}(t_k|t_{k-1})\right] & \end{aligned} \tag{4.37}$$

$$\begin{aligned}\Sigma(t_k|t_k) = \Sigma(t_k|t_{k-1}) & \\ - \Sigma(t_k|t_{k-1}) C_k^T \left[C_k \Sigma(t_k|t_{k-1}) C_k^T + R_k\right]^{-1} C_k \Sigma(t_k|t_{k-1}) & \end{aligned} \tag{4.38}$$

or, equivalently:

$$\Sigma^{-1}(t_k|t_k) = \Sigma^{-1}(t_k|t_{k-1}) + C_k^T R_k^{-1} C_k. \tag{4.39}$$

Continuous-time extrapolation from t_k to t_{k+1} with the initial conditions $\widehat{x}(t_k|t_k)$ and $\Sigma(t_k|t_k)$:

$$\dot{\widehat{x}}(t|t_k) = A(t)\widehat{x}(t|t_k) + B(t)\overline{u}(t) \tag{4.40}$$

$$\dot{\Sigma}(t|t_k) = A(t)\Sigma(t|t_k) + \Sigma(t|t_k)A^T(t) + B(t)Q(t)B^T(t)\ . \tag{4.41}$$

Assuming that the Kalman filter starts with an update at the initial time t_0, this Kalman filter is initialized at t_0 as follows:

$$\widehat{x}(t_0|t_{-1}) = x_0 \tag{4.42}$$

$$\Sigma(t_0|t_{-1}) = \Sigma_0\ . \tag{4.43}$$

Proof of Lemma 4.5 (update step):
A bias-free estimate $\widehat{x}_{k|k}$ can be obtained in the following form:

$$\widehat{x}_{k|k} = \widehat{x}_{k|k-1} + H_k[y_k - \overline{r}_k - C_k\widehat{x}_{k|k-1}]\ , \tag{4.44}$$

where H_k is an arbitrary observer gain matrix. Therefore, the estimation error $e = x - \widehat{x}$ changes as follows due to the processing of the measurement y_k:

$$\begin{aligned} e_{k|k} &= e_{k|k-1} - H_k[C_k e_{k|k-1} + r_k - \overline{r}_k] \\ &= [I - H_k C_k]e_{k|k-1} - H_k[r_k - \overline{r}_k]\ . \end{aligned} \tag{4.45}$$

By assumption, the white noise random vector r_k is independent of $e_{k|k-1}$. Therefore, the covariance matrix Σ of the estimation error e changes as a consequence to the processing of the new measurement y_k in the following way:

$$\Sigma_{k|k} = [I - H_k C_k]\Sigma_{k|k-1}[I - H_k C_k]^{\mathrm{T}} + H_k R_k H_k^{\mathrm{T}}\ . \tag{4.46}$$

The first derivative of $\Sigma_{k|k}$ with respect to H_k is

$$\frac{\partial \Sigma_{k|k}}{\partial H_k} = U[H_k C_k \Sigma_{k|k-1} C_k^{\mathrm{T}} + H_k R_k - \Sigma_{k|k-1} C_k^{\mathrm{T}}]T\ . \tag{4.47}$$

It vanishes if and only if the following optimal observer gain matrix

$$H_k = \Sigma_{k|k-1} C_k^{\mathrm{T}} [C_k \Sigma_{k|k-1} C_k^{\mathrm{T}} + R_k]^{-1} \tag{4.48}$$

is chosen. This choice yields the infimal error covariance matrix

$$\begin{aligned} \Sigma_{k|k} = \Sigma_{k|k-1} \\ - \Sigma_{k|k-1} C_k^T \left[C_k \Sigma_{k|k-1} C_k^T + R_k\right]^{-1} C_k \Sigma_{k|k-1}\ . \end{aligned} \tag{4.49}$$

Discrete-time version of the Kalman filter

In the discrete-time version of the problem, the system dynamics as well as the measurements are modeled in discrete-time. For this kind of problem, we consider the following discrete-time stochastic system:

$$x_{k+1} = F_k x_k + G_k \widetilde{v}_k \tag{4.50}$$
$$x_0 = \xi \tag{4.51}$$
$$y_k = C_k x_k + \widetilde{r}_k \tag{4.52}$$
$$\xi \ : \ \mathcal{N}(x_0, \Sigma_0) \tag{4.53}$$
$$\widetilde{v} \ : \ \mathcal{N}(\overline{u}_k, Q_k) \tag{4.54}$$
$$\mathrm{Cov}(\widetilde{v}_i, \widetilde{v}_j) = Q_i \delta_{ij} \tag{4.55}$$
$$\widetilde{r} \ : \ \mathcal{N}(\overline{r}_k, R_k) \tag{4.56}$$
$$\mathrm{Cov}(\widetilde{r}_i, \widetilde{r}_j) = R_i \delta_{ij} \tag{4.57}$$
$$\xi, \widetilde{v}, \widetilde{r} \ : \ \text{mutually independent.} \tag{4.58}$$

Of course, it is possible to transform the continuous-time version of the linear filtering problem into this discrete-time version. We have the following (approximate) "zero-order-hold-equivalence" correspondences to the continuous-time stochastic system:

$$F_k = \Phi(t_{k+1}, t_k) \qquad \text{transition matrix} \tag{4.59}$$
$$G_k = \int_{t_k}^{t_{k+1}} \Phi(t_{k+1}, t) B(t) dt \qquad \text{zero order hold equivalence} \tag{4.60}$$
$$C_k = C(t_k) \tag{4.61}$$
$$Q_k = \frac{Q(t_k)}{t_{k+1} - t_k} \tag{4.62}$$
$$R_k = \frac{R(t_k)}{\Delta t_k} \qquad \Delta t_k = \text{averaging time at time } t_k \tag{4.63}$$

Lemma 4.6. The Discrete-Time Kalman Filter
Update at time t_k:

$$\widehat{x}_{k|k} = \widehat{x}_{k|k-1} + \Sigma_{k|k-1} C_k^T \left[C_k \Sigma_{k|k-1} C_k^T + R_k \right]^{-1} [y_k - \overline{r}_k - C_k \widehat{x}_{k|k-1}] \tag{4.64}$$
$$\Sigma_{k|k} = \Sigma_{k|k-1} - \Sigma_{k|k-1} C_k^T \left[C_k \Sigma_{k|k-1} C_k^T + R_k \right]^{-1} C_k \Sigma_{k|k-1} \tag{4.65}$$

or, equivalently:

$$\Sigma_{k|k}^{-1} = \Sigma_{k-1|k-1}^{-1} + C_k^T R_k^{-1} C_k \tag{4.66}$$

Extrapolation from t_k to t_{k+1}:

$$\widehat{x}_{k+1|k} = F_k \widehat{x}_{k|k} + G_k \overline{u}_k \tag{4.67}$$

$$\Sigma_{k+1|k} = F_k \Sigma_{k|k} F_k^T + G_k Q_k G_k^T \tag{4.68}$$

Initialization at time t_0:

$$\widehat{x}_{0|-1} = x_0 \tag{4.69}$$

$$\Sigma_{0|-1} = \Sigma_0 \ . \tag{4.70}$$

Example 4.7. System of first order
Consider again the first-order stochastic system of Example 4.4 with the random initial condition $\xi : \mathcal{N}(x_0, \Sigma_0)$ and the uncorrelated white noises $\dot{v} : \mathcal{N}(0, Q)$ and $\dot{r} : \mathcal{N}(0, R)$:

$$\begin{aligned} \dot{x}(t) &= ax(t) + b[u(t) + \dot{v}(t)] \\ x(0) &= \xi \\ y(t) &= x(t) + \dot{r}(t)\,. \end{aligned}$$

In order to obtain reasonable results, the data are sampled with a constant sampling time $T \ll \tau = \frac{1}{|a|}$. — Find the discrete-time Kalman filter!

Using the zero order hold equivalence, the discrete-time version of the stochastic system is

$$\begin{aligned} x_{k+1} &= Fx_k + G[u_k + \widetilde{v}_k] \\ x_0 &= \xi \\ y_k &= x_k + \widetilde{r}_k \end{aligned}$$

with

$$\begin{aligned} F &= e^{aT} \\ G &= \frac{b}{a}\left(e^{aT} - 1\right) \\ \widetilde{v}_k &: \mathcal{N}(0, \widetilde{Q}) \qquad \widetilde{Q} = \frac{Q}{T} \\ \widetilde{r}_k &: \mathcal{N}(0, \widetilde{R})\,. \end{aligned}$$

The covariance matrix $\widetilde{R}$ of the discrete white noise signal $\widetilde{r}_k$ could be obtained from the intensity R of the continuous-time white noise $\dot{r}$ as $\widetilde{R} = \frac{R}{\Delta T}$, where ΔT is some averaging time of a continuous-time measurement. In practice, it is usually best to model the variance $\widetilde{R}$ of the discrete measurement error by considering the physics of the discrete measurement device.

The discrete-time Kalman filter is described by the following equations: Update at time t_k:

$$\widehat{x}_{k|k} = \widehat{x}_{k|k-1} + \frac{\Sigma_{k|k-1}}{\Sigma_{k|k-1} + \widetilde{R}}[y_k - \widehat{x}_{k|k-1}]$$
$$\Sigma_{k|k} = \Sigma_{k|k-1} - \frac{\Sigma_{k|k-1}^2}{\Sigma_{k|k-1} + \widetilde{R}}\,.$$

(Approximate) extrapolation from t_k to t_{k+1}:

$$\widehat{x}_{k+1|k} = F\widehat{x}_{k|k} + Gu_k$$
$$\Sigma_{k+1|k} = F^2\Sigma_{k|k} + G^2\widetilde{Q}\,.$$

Initialization at t_0:

$$\widehat{x}_{0|-1} = x_0$$
$$\Sigma_{0|-1} = \Sigma_0\,.$$

Notice that we have obtained a time-varying discrete Kalman filter. Since we have a time-invariant stochastic system, stationary noise processes, and a constant sampling time, we might as well use the time-invariant discrete Kalman filter (except perhaps during the start-up transient), where both $\Sigma_{k|k-1}$ and $\Sigma_{k|k}$ become constants. — The tedious algebra is left to the reader.

4.1.2 The Extended Kalman Filter

Continuous-time version of the extended Kalman filter

Obviously, there are many phenomena whose dynamics are not linear. Unfortunately, there is no closed theory for nonlinear problems as there is for linear problems. When the nonlinear model is too involved, a linearization of the problem is the only way to tackle the filtering problem. Since the linearized problem may not be a good approximation we stick to the following philosophy:

- Where it does not hurt: Analyze the nonlinear system.
- Where it hurts: use linearization.

We consider the following nonlinear stochastic system:

$$\dot{x}(t) = f(x(t), u(t), t) + B(t)\dot{v}(t) \tag{4.71}$$
$$x(t_0) = \xi \tag{4.72}$$
$$y(t) = g(x(t), t) + \dot{r}(t) \tag{4.73}$$
$$\xi \;:\; \mathcal{N}(x_0, \Sigma_0) \tag{4.74}$$
$$\dot{v} \;:\; \mathcal{N}(\overline{v}(t), Q(t)) \tag{4.75}$$
$$\dot{r} \;:\; \mathcal{N}(\overline{r}(t), R(t))\,. \tag{4.76}$$

The initial state ξ and the white noise processes $\dot{v}$ and $\dot{r}$ are assumed to be mutually independent. The above-mentioned pragmatic approach leads to the extended Kalman filter:

Lemma 4.8. The extended Kalman filter

Dynamics of the state estimation:

$$\dot{\widehat{x}}(t) = f(\widehat{x}(t), u(t), t) + B(t)\overline{v}(t) + H(t)[y(t) - \overline{r}(t) - g(\widehat{x}(t), t)] \quad (4.77)$$

$$\widehat{x}(t_0) = x_0 \quad (4.78)$$

with

$$H(t) = \Sigma(t)C^T(t)R^{-1}(t) \ . \quad (4.79)$$

The "error covariance" matrix $\Sigma(t)$ must be calculated in real-time using the folloing matrix Riccati differential equation:

$$\begin{aligned}\dot{\Sigma}(t) = {} & A(t)\Sigma(t) + \Sigma(t)A^T(t) \\ & - \Sigma(t)C^T(t)R^{-1}(t)C(t)\Sigma(t) + B(t)Q(t)B^T(t) \quad (4.80)\end{aligned}$$

$$\Sigma(t_0) = \Sigma_0 \ . \quad (4.81)$$

The matrices $A(t)$ and $C(t)$ correspond to the dynamics matrix and the output matrix, respectively, of the linearization of the nonlinear system around the estimated trajectory:

$$A(t) = \frac{\partial f(\widehat{x}(t), u(t), t)}{\partial x} \quad (4.82)$$

$$C(t) = \frac{\partial g(\widehat{x}(t), t)}{\partial x} \quad . \quad (4.83)$$

Because the dynamics of the "error covariance matrix" $\Sigma(t)$ correspond to the linearized system, we face the following problems:

- This filter is not optimal.
- The state estimate will be biased.
- The reference point for the linearization is questionable.
- The matrix $\Sigma(t)$ is only an approximation of the state error covariance matrix.
- The state vectors $x(t)$ und $\widehat{x}(t)$ are not Gaussian even if ξ, $\dot{v}$, and $\dot{r}$ are.

Continuous-time/discrete-time version of the extended Kalman filter

Again, the continous-time measurement corrupted with the white noise $\dot{r}$ with infinite covariance

$$y(t) = g(x(t), t) + \dot{r}(t) \tag{4.84}$$

must be replaced by an "averaged" dicrete-time measurement

$$y_k = g(x(t_k), t_k) + \widetilde{r}_k \tag{4.85}$$
$$\widetilde{r}_k \ : \ \mathcal{N}(\overline{r}_k, R_k) \tag{4.86}$$
$$\mathrm{Cov}(\widetilde{r}_i, \widetilde{r}_j) = R_i \delta_{ij} \ . \tag{4.87}$$

This version of the extended Kalman filter works as follows:

Lemma 4.9. The continuous-time/discrete-time extended Kalman filter
Update at time t_k:

$$\begin{aligned}\widehat{x}(t_k|t_k) &= \widehat{x}(t_k|t_{k-1}) \\ &\quad + \Sigma(t_k|t_{k-1}) C_k^T \left[C_k \Sigma(t_k|t_{k-1}) C_k^T + R_k\right]^{-1} \\ &\qquad \times [y_k - \overline{r}_k - g(\widehat{x}(t_k|t_{k-1}), t_k)] \end{aligned} \tag{4.88}$$
$$\Sigma(t_k|t_k) = \Sigma(t_k|t_{k-1}) \tag{4.89}$$
$$- \Sigma(t_k|t_{k-1}) C_k^T \left[C_k \Sigma(t_k|t_{k-1}) C_k^T + R_k\right]^{-1} C_k \Sigma(t_k|t_{k-1}) \tag{4.90}$$

Extrapolation between t_k and t_{k+1}:

$$\dot{\widehat{x}}(t|t_k) = f(\widehat{x}(t|t_k), u(t), t) + B(t)\overline{v}(t) \tag{4.91}$$
$$\dot{\Sigma}(t|t_k) = A(t)\Sigma(t|t_k) + \Sigma(t|t_k)A^T(t) + B(t)Q(t)B^T(t) \tag{4.92}$$

Again, the matrices $A(t)$ and C_k are the dynamics matrix and the output matrix, respectively, of the nonlinear system linearized around the estimated trajectory:

$$A(t) = \frac{\partial f(\widehat{x}(t), u(t), t)}{\partial x} \tag{4.93}$$
$$C_k = \frac{\partial g(\widehat{x}(t_k|t_{k-1}), t_k)}{\partial x} \tag{4.94}$$

Initialization at t_0:

$$\widehat{x}(t_0|t_{-1}) = x_0 \tag{4.95}$$
$$\Sigma(t_0|t_{-1}) = \Sigma_0 \ . \tag{4.96}$$

Discrete-time version of the extended Kalman filter

In general, it will not be possible to integrate the differential equations of the nonlinear dynamic system analytically. Therefore, there is no discrete-time version of the extended Kalman filter which would be analogous to the discrete-time Kalman filter of Lemma 4.6. Rather, the differential equations for the state estimate $\widehat{x}(t)$ and the "error covariance matrix" $\Sigma(t)$ have to be numerically integrated over the time interval $[t_k, t_{k+1}]$ between two subsequent measurements.

4.2 Nonlinear Filtering

Continuous-time version

The following stochastic nonlinear dynamic system is considered:

$$\begin{aligned}
\dot{x}(t) &= f(x(t), u(t), t) + B(x(t), t)\dot{v}(t) \\
x(t_0) &= \xi \\
y(t) &= g(x(t), t) + \dot{r}(t) \\
\frac{\partial g(x,t)}{\partial x} &: \text{ has full rank for all } x \text{ and } t \\
\xi &: \mathcal{N}(x_0, \Sigma_0) \\
\dot{v} &: \mathcal{N}(\overline{v}(t), Q(x(t), t)) \\
\dot{r} &: \mathcal{N}(\overline{r}(t), R(t)) \,.
\end{aligned}$$

The initial state ξ and the white noise processes $\dot{v}$ and $\dot{r}$ are assumed to be mutually independent. We use the following

Philosophy:

- Optimal filtering implies making full usage of information, i.e., probability density functions must be calculated in continuous-time.
- In general, state estimates $\widehat{x}(.)$ will only be needed at discrete times t_k.
- For a given probability density function $p_{\widehat{x}(t_k)}$ for $\widehat{x}(t_k)$, there are various methods for extracting a suitable estimate $\widehat{x}(t)$.

The continuous-time version of optimal filtering is of interest only if all of the relevant equations can be solved analytically. — In general, we will have to live with a continuous-time/discrete-time version of optimal nonlinear filtering.

Continuous-time/discrete-time version

Again, the continous-time measurement corrupted with the white noise $\dot{r}$ with infinite covariance

$$y(t) = g(x(t), t) + \dot{r}(t)$$

must be replaced by an "averaged" dicrete-time measurement

$$\begin{aligned} y_k &= g(x(t_k), t_k) + \widetilde{r}_k \\ \widetilde{r}_k &: \mathcal{N}(\overline{r}_k, R_k) \\ \mathrm{Cov}(\widetilde{r}_i, \widetilde{r}_j) &= R_i \delta_{ij} \ . \end{aligned}$$

The following continuous-time/disrete-time scheme for optimal nonlinear filtering can be guessed from the (extended) Kalman filter:

Theorem 4.10. The optimal nonlinear filter

Update at time t_k:
Calculate the conditional probability density function of $x(t_k|t_k)$ using the multivariate probability density function of $y(t_k|t_{k-1})$ and $x(t_k|t_{k-1})$ and the probability density function of $y(t_k|t_{k-1})$ at the measured value $y(t_k)$:

$$\begin{aligned} p(x(t_k|t_k)) &= \frac{p(y(t_k|t_{k-1}), x(t_k|t_{k-1}))}{p(y(t_k|t_{k-1}))_{|y(t_k)}} \\ &= \frac{p(y(t_k|t_{k-1})|x(t_k|t_{k-1}))}{p(x(t_k|t_{k-1}))} p(y(t_k|t_{k-1}))_{|y(t_k)} \ . \end{aligned}$$

Please note:

- $p(x(t_k|t_{k-1}))$: *solution of the corresponding Fokker-Planck partial differential equation*
- $p(y_k|x_k) \leftrightarrow \mathcal{N}(g(x_k, t_k) + \overline{r}_k, R_k)$
- $p(y(t_k|t_{k-1}))$: *to be computed from $p(x(t_k|t_{k-1}))$ and $p(\widetilde{r}_k)$ using the general relation*

$$p(y) = \int \ldots \int p(y|x) p(x) \, dx \ .$$

Extrapolation between t_k and t_{k+1}:
Calculate the probability density function $p(x(t|t_k))$ of $x(t|t_k)$ by solving the following Fokker-Planck partial differential equation over the time interval $[t_k, t_{k+1}]$:

$$\begin{aligned} &\frac{\partial p(x,t)}{\partial t} + \sum_i \frac{\partial}{\partial x_i} \Big\{ \Big[f_i(x, u(t), t) + B_i(x,t) \overline{v}(t) \Big] p(x,t) \Big\} \\ &- \frac{1}{2} \sum_i \sum_j \frac{\partial^2}{\partial x_i \partial x_j} \Big\{ \Big[B(x,t) Q(x,t) B^T(x,t) \Big]_{ij} p(x,t) \Big\} = 0 \ . \end{aligned}$$

Initialization at time t_0:

$$p(x(t_0|t_{-1})) = p(\xi) \ .$$

4.3 Kalman Filter and Parameter Identification

4.3.1 Introduction

In order to use stochastic differential equations as a tool to model financial prices or technical applications, one important step is to identify the relevant parameters of one's models. In contrast to engineering applications, first-principle models hardly exist for financial problems. Thus, one has to resort to quantitative techniques to identify the parameters of the chosen models. In this chapter, we use the Kalman filter in order to derive a *maximum likelihood estimator*.

4.3.2 Kalman Filter Equations

For ease of reference, the equations of the discrete-time Kalman filter are recapitulated here. In addition, we show some results of probability theory which are useful in the derivation of the maximum likelihood estimator.

We consider the discrete-time stochastic dynamic system

$$x_{k+1} = F_k x_k + G_k \widetilde{v}_k \tag{4.97}$$
$$x_0 = \xi \tag{4.98}$$
$$y_k = C_k x_k + \widetilde{r}_k \tag{4.99}$$

with the stochastic properties

$$\xi \ : \ \mathcal{N}(x_0, \Sigma_0) \tag{4.100}$$
$$\widetilde{v}_k \ : \ \mathcal{N}(\overline{u}_k, Q_k) \tag{4.101}$$
$$\mathrm{Cov}(\widetilde{v}_i, \widetilde{v}_j) = Q_i \delta_{ij} \tag{4.102}$$
$$\widetilde{r}_k \ : \ \mathcal{N}(\overline{r}_k, R_k) \tag{4.103}$$
$$\mathrm{Cov}(\widetilde{r}_i, \widetilde{r}_j) = R_i \delta_{ij} \ . \tag{4.104}$$

We assume that ξ, $\widetilde{v}_{(.)}$, and $\widetilde{r}_{(.)}$ are mutually independent.
The equations of the discrete-time Kalman filter are:

Update at time t_k due to the measurement y_k:

$$\begin{aligned} e_{k|k-1} &= [y_k - \overline{r}_k - C_k \widehat{x}_{k|k-1}] \\ \widehat{x}_{k|k} &= \widehat{x}_{k|k-1} + \Sigma_{k|k-1} C_k^T [C_k \Sigma_{k|k-1} C_k^T + R_k]^{-1} e_{k|k-1} \\ \Sigma_{k|k} &= \Sigma_{k|k-1} - \Sigma_{k|k-1} C_k^T [C_k \Sigma_{k|k-1} C_k^T + R_k]^{-1} C_k \Sigma_{k|k-1} \ , \end{aligned}$$

where $e_{k|k-1}$ denotes the unbiased prediction error or residual at the output, conditioned on the measurements observed up to and including time t_{k-1}.

Extrapolation from time t_k to time t_{k+1}:

$$\begin{aligned}\widehat{x}_{k+1|k} &= F_k\widehat{x}_{k|k} + G_k\overline{u}_k\\ \Sigma_{k+1|k} &= F_k\Sigma_{k|k}F_k^T + G_kQ_kG_k^T\end{aligned}$$

Initialization at t_0:

$$\begin{aligned}\widehat{x}_0 &= x_0\\ \Sigma_{0|-1} &= \Sigma_0\,.\end{aligned}$$

Note that due to the mutual independence of ξ, $\widetilde{v}_{(.)}$, and $\widetilde{r}_{(.)}$, the unbiased conditional prediction error

$$e_{k|k-1} = y_k - \overline{r}_k - C_k\widehat{x}_{k|k-1} \tag{4.105}$$

has the covariance matrix

$$\Lambda_{k|k-1} = C_k\Sigma_{k|k-1}C_k^T + R_k\,. \tag{4.106}$$

This result is important for building the log-likelihood estimator.

4.3.3 Parameter Estimation

In this section, we denote by $Y = \{Y_0, Y_1, \ldots, Y_{N-1}, Y_N\}$ the discrete-time ramdom process which corresponds to the output of the stochastic system (4.97)–(4.104) for the time interval from t_0 through t_N. On the other side, we denote by $y = \{y_0, y_1, \ldots, y_{N-1}, y_N\}$ an individual sample path of Y which is obtained from the measurements at the times t_0 through t_N in a single experiment. — This should help the reader distinguish between a priori information which can be obtained from Y and a posteriori or conditional information based on the sample path y.

In the scenario of this section, some or all of the elements of the matrices F_k, G_k, Σ_0, Q_k, and R_k are not perfectly known. But they are parametrized by a "completely" unknown parameter vector $\psi \in \Psi \subseteq \mathbb{R}^d$.

Obviously, the multivariate probability density function of the random process Y is parametrized by ψ and can be denoted as follows:

$$p(Y,\psi)(.)\;.$$

Given the sample path y of the measurements, we introduce the so-called likelihood function

$$l(\psi|y) = p(Y,\psi)(y) = p(Y,\psi \mid y_0, y_1, \ldots, y_{N-1}, y_N)\,.$$

Notice that the likelihood function l is a function of ψ alone because for this sample path, the values y_0, …, y_N are fixed. This is explicitly denoted by

"given that ..." in $l(.)$ and on the right-hand side. This is standard in the context of conditional probability density functions.

Using Bayes' rule for probability density functions repetitively, we can write

$$\begin{aligned} l(\psi|y) &= p(Y,\psi)(y) \\ &= p(Y_N,\psi \mid y_0,\ldots,y_{N-1})(y_N) \\ &= p(Y_N,\psi \mid y_0,\ldots,y_{N-1})(y_N)\cdot p(Y_{N-1},\psi \mid y_0,\ldots,y_{N-2})(y_{N-1}) \\ &\vdots \\ &= p(Y_N,\psi \mid y_0,\ldots,y_{N-1})(y_N)\cdot p(Y_{N-1},\psi \mid y_0,\ldots,y_{N-2})(y_{N-1}) \\ &\quad \cdot \ldots \cdot p(Y_k,\psi \mid y_0,\ldots,y_{k-1})(y_k)\cdot \ldots \cdot p(Y_1,\psi \mid y_0)(y_1)\cdot p(Y_0,\psi)(y_0)\,. \end{aligned}$$

Since the stochastic processes X and Y are Markov processes, conditioning the general term $p(Y_k,\psi \mid y_0,\ldots,y_{k-1})(y_k)$ on y_0 through y_{k-1} amounts to the same as conditioning on the most recent value y_{k-1} only. Thus, we get the considerably simpler expression

$$\begin{aligned} l(\psi|y) = p(Y_N,\psi \mid y_{N-1})(y_N)\cdot p(Y_{N-1},\psi \mid y_{N-2})(y_{N-1})\cdot \ldots \\ \cdot p(Y_k,\psi \mid y_{k-1})(y_k)\cdot \ldots \cdot p(Y_1,\psi \mid y_0)(y_1)\cdot p(Y_0,\psi)(y_0)\,. \end{aligned} \tag{4.107}$$

Given the measurements $y_0,\ldots, y_N$, we want to choose the "most likely" value for the unknown parameter ψ. By this, we mean that we want to find the value of ψ which maximizes the likelihood function $l(\psi|y)$.

Equation (4.105) shows that the *conditional* density functions $p(Y_k,\psi|y_{k-1})$ of the measurement at time t_k and $p(E_k,\psi|y_{k-1})$ of the corresponding residual only differ by "bias" terms. Therefore, in Equation (4.107), we can replace the former by the latter.

Since all of the distributions are Gaussian, it is more practical to consider the so-called "log-likelihood function" $L(\psi|y) = \ln(l(\psi|y))$. This yields the expression

$$L(\psi|y) = -\frac{1}{2}\sum_{k=0}^{N}\Big(p\ln(2\pi) + \ln(\det \Lambda_{k|k-1}) + e_{k|k-1}^T \Lambda_{k|k-1}^{-1} e_{k|k-1}\Big)\,. \tag{4.108}$$

Since the logarithm is a monotonic function, the desired maximum likelihood estimate $\widehat{\psi}_{ML}(y)$ for the unknown parameter $\psi \in \Psi \subseteq \mathbb{R}^d$ is:

$$\widehat{\psi}_{ML}(y) = \arg\max_{\psi\in\Psi} L(\psi|y)\,. \tag{4.109}$$

In order to estimate the unknown parameters from the log-likelihood function in (4.108), we use an appropriate optimization routine which maximizes $L(\psi|y)$ with respect to ψ. Given the simplicity of the innovation form of the log-likelihood function, we can find the values for the conditional expectation

of the prediction error and the conditional covariance matrix for every given parameter vector ψ using the Kalman filter algorithm. Hence, we are able to calculate the value of the conditional log-likelihood function numerically.

The flow chart in Figure 4.1 gives a graphical illustration of the Kalman filter parameter estimation algorithm. The recursion starts with a feasible choice of the parameter vector ψ and a state vector x_0 with covariance matrix Σ_0. Thereupon, we use the prediction equations to calculate the a priori estimates, then use the update equations to calculate the a posteriori estimates. For each time step, the prediction error and the covariance matrix are saved. This loop is repeated until we reach the final time T after N steps. Finally, with the obtained time-series of the state variables we can evaluate the likelihood function to choose a better estimate of ψ. The Kalman filter recursion is started again, and the likelihood function is evaluated again and so on, until a reasonable stopping criterion is satisfied.

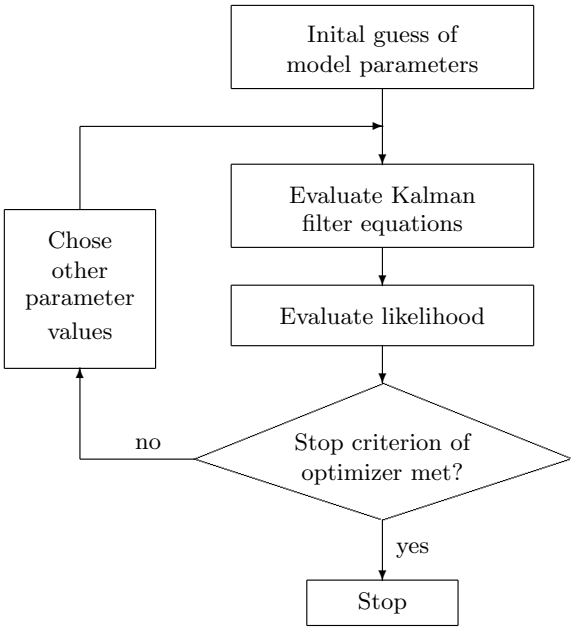

Fig. 4.1. Flow Chart of Kalman Filter Parameter Estimation

In the case of implementing extended Kalman filter algorithms where we use approximations of the conditional normal distribution, the log-likelihood function (4.108) can be used to yield a quasi maximum likelihood parameter estimate.

4.3.4 Numerical Implementation

In the case where the parameter space is unrestricted, the estimation includes an unconstrained optimization. For this case, we briefly discuss a numerical procedure to calculate a better estimate of ψ. In the numerical implementation to obtain the optimal parameter estimates, we actually minimize the negative log-likelihood function of Equation (4.108).

$$\widehat{\psi}_{ML}(y) = \arg\min_{\psi \in \Psi} -L(\psi|y)$$

The necessary condition for a minimum is

$$\frac{\partial L}{\partial \psi}(\widehat{\psi}_{ML}|y) = 0$$

assuming that the relevant partial derivative exists. A sufficient condition is that the Hessian matrix of $= -L$ exists and is positive-definite, i.e.,

$$-\frac{\partial^2 L}{\partial \psi^2}(\widehat{\psi}_{ML}|y) > 0\,.$$

In order to compute the minimum numerically, we choose a quasi-Newton method. The iteration rule for Newton's method for unconstrained optimization is given by the parameter estimates at the i^{th} iteration step by:

$$\psi_{i+1} = \psi_i + s_i\,,$$

where the variable s_i is found by solving

$$\frac{\partial^2 L}{\partial \psi^2}(\psi_i|y) \cdot s_i = -\frac{\partial L}{\partial \psi}(\psi_i|y)\,.$$

In order to calculate $\frac{\partial L}{\partial \psi}(\psi_i|y)$ and $\frac{\partial^2 L}{\partial \psi^2}(\psi_i|y)$ we use numerical approximations, such as first and second order differences. Since the log-likelihood function assumes a global maximum, the Hessian is negative-definite and thus the quasi-Newton method converges.

Notes and Comments

For more details on this chapter, the following books are recommended: [16], [3], [2], [21], [32], [13], [22], [12], [19], [1], and [18].

5

Optimal Control

As in all optimal control problems,
the GIGO principle applies.
Hans P. Geering

5.1 Deterministic Optimal Control

We begin the section on optimal control with a repetition of the deterministic case. We do this since the methods and results of the deterministic case can be regarded as a special case of the stochastic case. In addition, it is essential to see the differences and similarities in the methods as well as the results of the deterministic and the stochastic cases.

In the deterministic optimal control problem, we regard a deterministic dynamic system, a control, and a cost functional which is to be minimized. In the optimal control problem, the control strategy is sought that effectively minimizes the cost functional under the given dynamics. Two approaches are given here: Pontryagin's minimum principle and Bellman's dynamic programming method leading to the well known Hamilton-Jacobi-Bellman equation.

The solution procedure for Pontryagin's minimum principle involves solving a two-point boundary problem. The resulting optimal control strategy is then the optimal feed-forward solution. The optimal feed-back strategy is obtained by solving the Hamilton-Jacobi-Bellman partial differential equation. For an in depth treatment of the matter of optimal control we refer to [8], [15], or [**?**].

5.1.1 Deterministic Optimal Control Problems

Consider the following dynamical system with the state vector $x(t) \in \mathbb{R}^n$ and the control vector $u(t) \in \mathbb{R}^m$:

$$\dot{x}(t) = f(x(t), u(t), t)\,.$$

Its initial state at the fixed initial time 0 is given:

$$x(0) = x_0\,.$$

The permissible controls over the fixed time interval $[0, T]$ satisfy the following condition:

$$u(t) \in \mathcal{U} \quad \text{for all } t \in [0, T],$$

where $\mathcal{U}$ is a time-invariant, closed, and convex subset of the control space $\mathbb{R}^m$:

$$\mathcal{U} \subseteq \mathbb{R}^m .$$

Furthermore, consider a cost functional of the following form:

$$J = K(x(T)) + \int_0^T L(x(t), u(t), t)\, dt .$$

This cost functional should either be minimized or maximized, depending upon the problem at hand. Consequently, there are two alternative formulations of the optimal control problem.

The Minimization problem:
Find the control trajectory $u^* : [0, T] \to \mathcal{U} \subseteq \mathbb{R}^m$ generating the state trajectory $x^* : [0, T] \to \mathbb{R}^n$ such that the cost functional J is minimized.

The Maximization problem:
Find the control trajectory $u^* : [0, T] \to \Omega \subseteq \mathbb{R}^m$ generating the state trajectory $x^* : [0, T] \to \mathbb{R}^n$ such that the cost functional J is maximized.

Definition 5.1. Hamiltonian function
The Hamiltonian function $H : \mathbb{R}^n \times \mathcal{U} \times \mathbb{R}^n \times [0, T]$ associated with a regular optimal control problem is:

$$H(x(t), u(t), p(t), t) = L(x(t), u(t), t) + p^T(t) f(x(t), u(t), t) \ ,$$

where $p(t) \in \mathbb{R}^n$ is the so-called costate vector.

5.1.2 Necessary Conditions for Optimality

The Russian mathematician Pontryagin has found the following necessary conditions for the optimality of a solution:

Theorem 5.2. Pontryagin
If $u^* : [0, T] \to \mathcal{U}$ is an optimal control trajectory, the following conditions are satisfied:

a) Optimal state trajectory:

$$\begin{aligned} \dot{x}^*(t) &= \nabla_p H|_* \\ &= f(x^*(t), u^*(t), t) \quad \textit{for } t \in [0, T] \\ x^*(0) &= x_0 \end{aligned}$$

b) Optimal costate trajectory: There exists an optimal costate trajectory satisfying

$$\begin{aligned}\dot{p}^*(t) &= -\nabla_x H|_* \\ &= -\nabla_x L(x^*(t), u^*(t), t) - f_x^T(x^*(t), u^*(t), t)p^*(t) \quad \textit{for } t \in [0,T] \\ p^*(T) &= \nabla_x K(x^*(T))\,.\end{aligned}$$

c) Global static optimization of the Hamiltonian function:
For the minimization problem:
For all $t \in [0,T]$, the Hamiltonian is globally minimized w.r. to u, i.e.,

$$H(x^*(t), u^*(t), p^*(t), t) \leq H(x^*(t), u, p^*(t), t) \quad \textit{for all } u \in \mathcal{U}\ .$$

For the maximization problem:
For all $t \in [0,T]$, the Hamiltonian is globally maximized w.r. to u, i.e.,

$$H(x^*(t), u^*(t), p^*(t), t) \geq H(x^*(t), u, p^*(t), t) \quad \textit{for all } u \in \mathcal{U}\ .$$

5.1.3 Example: The LQ-Regulator Problem

For the linear time-varying system

$$\dot{x}(t) = A(t)x(t) + B(t)u(t)$$

with the initial state

$$x(0) = x_0$$

find the unconstrained optimal control $u : [0,T] \to \mathbb{R}^m$ such that the quadratic cost functional

$$J = \frac{1}{2}x^T(T)Fx(T) + \int_0^T \frac{1}{2}\Big(x^T(t)Q(t)x(t) + u^T(t)R(t)u(t)\Big)dt$$

is minimized.

Here, the penalty matrix $R(t)$ is symmetric and positive-definite, and the penalty matrices F and $Q(t)$ are symmetric and positive-semidefinite.

Analysis of the necessary conditions for optimality:

Hamiltonian function:

$$\begin{aligned}H(x(t), u(t), p(t), t) = &\frac{1}{2}x^T(t)Q(t)x(t) + \frac{1}{2}u^T(t)R(t)u(t) \\ &+p^T(t)A(t)x(t) + p^T(t)B(t)u(t)\end{aligned}$$

Pontryagin's necessary conditions for optimality:

$$\begin{aligned}
\dot{x}^*(t) &= A(t)x^*(t) + B(t)u^*(t) \\
x^*(0) &= x_0 \\
\dot{p}^*(t) &= -Q(t)x^*(t) - A^T(t)p^*(t) \\
p^*(T) &= Fx^*(T) \\
u^*(t) &= \arg\min_{u\in\mathbb{R}^m} \left(\frac{1}{2}u^T R(t)u + p^*(t)B(t)u\right)
\end{aligned}$$

H-minimizing control:

$$u^*(t) = -R^{-1}(t)B^T(t)p^*(t)$$

Plugging the H-minimizing control into the differential equations leads to the following linear two-point-boundary-value problem:

$$\begin{aligned}
\dot{x}^*(t) &= A(t)x^*(t) - B(t)R^{-1}(t)B^T(t)p^*(t) \\
\dot{p}^*(t) &= -Q(t)x^*(t) - A^T(t)p^*(t) \\
x^*(0) &= x_0 \\
p^*(T) &= Fx^*(T)\,.
\end{aligned}$$

Considering that $p^*(T)$ is linear in $x^*(T)$ and that the linear differential equations are homogeneous leads to the educated guess that $p^*(t)$ is linear in $x^*(t)$ at all times, i.e.,

$$p^*(t) = K(t)x^*(t)$$

with a suitable n by n matrix function $K(t)$.

This leads to the following linear state feedback control:

$$u^*(t) = -R^{-1}(t)B^T(t)K(t)x^*(t)\,.$$

Exploiting the two-point-boundary-value problem and the proposed linear relation leads to the following matrix Riccati differential equation for $K(t)$ with a boundary condition at the final time T:

$$\begin{aligned}
\dot{K}(t) &= -A^T(t)K(t) - K(t)A(t) + K(t)B(t)R^{-1}(t)B^T(t)K(t) - Q(t) \\
K(T) &= F\ .
\end{aligned}$$

5.1.4 Deterministic Hamilton-Jacobi-Bellman Theory

In this section, we give a short introduction to the deteministic Hamilton-Jacobi-Bellman theory.

Consider again the optimal control problem formulated in Section 5.1.1:

$$\begin{aligned}
\dot{x}(t) &= f(x(t), u(t), t) \qquad u(t) \in \mathcal{U} \subseteq \mathbb{R}^m \\
x(0) &= x_0 \\
J &= K(x(T)) + \int_0^T L(x(t), u(t), t)\, dt \ :\ \min!
\end{aligned}$$

Assumptions:

i) Let the Hamiltonian function $H(x,u,p,t) = L(x,u,t) + p^T f(x,u,t)$ as a function of u, $u \in \mathcal{U}$, have a unique absolute minimum for all $p \in \mathbb{R}^n$ and all $(x,t) \in X \subseteq \mathbb{R}^n \times [0,T]$ at

$$\widetilde{u}(x,p,t) \quad .$$

ii) Let $\mathcal{J}(x,t) : X \to \mathbb{R}$ be a continuously differentiable function satisfying the following partial differential equation

$$\frac{\partial \mathcal{J}(x,t)}{\partial t} + H\Big[x, \widetilde{u}\big(x, \nabla_x \mathcal{J}(x,t), t\big), \nabla_x \mathcal{J}(x,t), t\Big] = 0$$

and the following boundary condition at the final time T:

$$\mathcal{J}(x,T) = K(x) \qquad \text{for all } (x,T) \in X \ .$$

Theorem 5.3. Hamilton-Jacobi-Bellman
If the above-mentioned assumptions are satisfied and if the control trajectory $\widehat{u} : [0,T] \to \mathcal{U}$ *generating the state trajectory* $\widehat{x} : [0,T] \to \mathbb{R}^n$ *satisfies the condition*

$$\widehat{u}(t) = \widetilde{u}\big(\widehat{x}(t), \nabla_x \mathcal{J}(\widehat{x}(t),t), t\big) \qquad \textit{for all } t \in [0,T] \ ,$$

then the solution $\big(\widehat{u}(.), \widehat{x}(.)\big)$ *is optimal with respect to all of the admissible trajectories* $x(.)$ *which do not leave* X*.*

Remarks:

i) $\mathcal{J}(x,t)$ is the optimal-cost-to-go function.
ii) The above statement of the theorem can be adapted to the problem of maximizing the cost functional J in the obvious way.
iii) If $X = \mathbb{R}^n \times [0,T]$, then the theorem states both necessary *and* sufficient conditions for the unique globally optimal solution of the control problem.

5.1.5 Example: The LQ-Regulator Problem

Consider again the LQ regulator problem:

$$\dot{x}(t) = A(t)x(t) + B(t)u(t)$$
$$x(0) = x_0$$
$$J = \frac{1}{2}x^T(T)Fx(T) + \int_0^T \frac{1}{2}\Big(x^T(t)Q(t)x(t) + u^T(t)R(t)u(t)\Big)dt : \min!$$

Quite generally speaking: Since this is an infinite-dimensional least-squares problem with linear side constraints, the solution is linear and the optimal value of the cost J is quadratic (in x_0).

Therefore, the optimal-cost-to-go function $\mathcal{J}(x,t)$ must be of the form

$$\mathcal{J} = \frac{1}{2}x^T K(t)x$$

with a suitable symmetric and positive-(semi)definite n by n matrix $K(t)$, defined for $t \in [0,T]$.

Analysis:

$$\begin{aligned}\frac{\partial \mathcal{J}(x,t)}{\partial t} &= \frac{1}{2}x^T \dot{K}(t)x \\ \nabla_x \mathcal{J}(x,t) &= K(t)x \\ \widetilde{u}(x,p,t) &= -R^{-1}(t)B^T(t)p = -R^{-1}(t)B^T(t)K(t)x\end{aligned}$$

HBJ partial differential equation:

$$\frac{1}{2}x^T\Big\{\dot{K}(t) + Q + A^T K + KA - KBR^{-1}B^T K\Big\}x = 0$$

Boundary condition:

$$\mathcal{J}(x,T) = \frac{1}{2}x^T K(T)x = \frac{1}{2}x^T Fx$$

Thus, the result of Section 5.1.3 is reestablished.

5.2 Stochastic Optimal Control

After having recalled the deterministic optimal control problem, we shall now generalize the deterministic system dynamics to the stochastic case by including diffusion models. As in the deterministic case, we are interested in finding an optimal control that minimizes the *expected* cost functional subjected to the stochastic dynamic system. The control may independently influence both the deterministic and the stochastic part of the system dynamics. The range of stochastic optimal control problems cover a great variety of fields, such as economics, physics, and management systems among others.

The topic of stochastic optimal control is well covered in the literature, among others in [6], [15], and [37].

5.2.1 Stochastic Optimal Control Problems

Consider the following dynamical system with state vector $x(t) \in \mathbb{R}^n$ and control vector $u(t) \in \mathbb{R}^m$:

$$\begin{aligned}dx(t) &= f(x(t),u(t),t)dt + g(x(t),u(t),t)dW(t) \\ x(0) &= x_0\,.\end{aligned} \tag{5.1}$$

The functions

$$f : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}^n$$
$$g : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}^{n \times k}$$

are given. Let $W(t)$ be a k-dimensional Brownian motion which is given on the probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, \mathbf{P})$ on which the usual conditions apply. Furthermore, over the fixed time interval $[0, T]$, we restrict the control vector $u(t)$ such that $u(t) \in \mathcal{U}$ for all $t \in [0, T]$. The set of permissible controls $\mathcal{U}$ is a time-invariant, closed, and convex subset of the control space $\mathbb{R}^m$.

We further consider the cost functional

$$J(u) = \mathrm{E}\left[K(x(T)) + \int_0^T L(x(t), u(t), t)\, dt \right] \tag{5.2}$$

with the given scalar functions $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}$ and $K : \mathbb{R}^n \to \mathbb{R}$.

Problem 5.4. Maximize (5.2) by finding $u^*(t) : [0, T] \to \mathcal{U} \subseteq \mathbb{R}^m$ satisfying

$$J(u^*) = \max_{u:[0,T] \to \mathcal{U}} J(u). \tag{5.3}$$

The corresponding optimal state trajectory $x^* : [0, T] \to \mathbb{R}^n$ and the optimal control $u^* : [0, T] \to \mathcal{U}$ form the optimal pair (x^*, u^*). We define the optimal *cost-to-go* function $J(t, x)$ as

$$J(0, x_0) = J(u^*) = \max_{u:[t,T] \to \mathcal{U}} J(u)\,. \tag{5.4}$$

Remark: Obviously, there is another optimal control problem, where the cost functional is to be minimized.

5.2.2 Stochastic Hamilton-Jacobi-Bellman Equation

Consider the optimal control problem stated in Section 5.2.1, equations (5.1)–(5.4) and the Problem 5.4.

The steps for finding a solution to the stochastic optimal control problem follow the approach by R. Bellmann called *dynamic programming*. Essentially, the idea is that an optimal path has the property, that whatever the initial conditions and the control values over a certain initial period are, the control variables over the remaining period must be optimal for the remaining problem.

This means, the problem can be divided into two parts, where we assume that from time $t + \Delta t$ to time T we have computed the optimal solution and we are still searching for the solution from time t to time $t + \Delta t$. It is sufficient to search for the solution of the first part, since the second part is already optimal:

$$u(t) = \begin{cases} u(s) & \text{for } t \le s \le t + \Delta t \\ u^*(s) & \text{for } t + \Delta t < s \le T \end{cases} \tag{5.5}$$

The cost-to-go function can now be split into the two parts and we obtain

$$\begin{aligned} J(x,t) = \max_{u:[t,T]\to\mathcal{U}} \mathrm{E}\Bigg[& \int_t^{t+\Delta t} L(x(t),u(t),t)\,dt \\ & \underbrace{+ K(x(T)) + \int_{t+\Delta t}^{T} L(x(t),u(t),t)\,dt\Bigg]}_{J(x,t+\Delta t)} \end{aligned} \tag{5.6}$$

We assume, that the control function $u(t)$ in $t_0 + \Delta t \le t \le T$ is optimal for the problem beginning at $t = t_0 + \Delta t$ so that we can write the second integral as $J(x, t + \Delta t)$. We then substitute this function by taking the total differential and integrating it. We must be aware that x is a stochastic process and we therefore need to apply the Itô calculus. This results in the following derivation for $J(x, t + \Delta t)$

$$J(x,t+\Delta t) = J(x,t) + \int_t^{t+\Delta t} \left(\frac{\partial J(x,t)}{\partial t} + \mathcal{A}J(x,t) \right) dt \tag{5.7}$$

$$+ \int_t^{t+\Delta t} J_x(x,x) g(x,u,t) dW, \tag{5.8}$$

where we use the stochastic differential operator

$$\mathcal{A}(\cdot) = \frac{\partial(\cdot)}{\partial x} f(x,u,t) + \frac{1}{2}\mathrm{tr}\Big\{ g(x,u,t) g^T(x,u,t) \frac{\partial^2(\cdot)}{\partial x^2} \Big\}$$

Now, we plug $J(x, t + \Delta t)$ into (5.6), and we obtain

$$\begin{aligned} J(x,t) &= \max_{u:[t,t+\Delta t]\to\mathcal{U}} \mathrm{E}\left[J(x,t+\Delta t) + \int_t^{t+\Delta t} L(x(t),u(t),t)\,dt \right] \\ &= \max_{u:[t,t+\Delta t]\to\mathcal{U}} \mathrm{E}\Bigg[J(x,t) + \int_t^{t+\Delta t} L(x,u,t)\,dt \\ &\quad + \int_t^{t+\Delta t} \left(\frac{\partial J(t,x)}{\partial t} + \mathcal{A}J(x,t) \right) dt + \int_t^{t+\Delta t} J_x(x,t) g(x,u,t) dW \Bigg] \\ &= \max_{u:[t,t+\Delta t]\to\mathcal{U}} \mathrm{E}\Bigg[J(x,t) + \int_t^{t+\Delta t} L(x,u,t) dt \\ &\quad + \int_t^{t+\Delta t} \left(\frac{\partial J(x,t)}{\partial t} + \mathcal{A}J(x,t) \right) dt \Bigg], \end{aligned} \tag{5.9}$$

We now take the expectation which removes the stochastic integral and subtract $J(t,x)$ on both sides (which we can do as this is a deterministic function independent of u and therefore not affected by the maximization over u) leading to:

$$0 = \max_{u:[t,t+\Delta t]\to\mathcal{U}} \mathrm{E}\left[\int_t^{t+\Delta t} \left(L(x,u,t) + \frac{\partial J(x,t)}{\partial t} + \mathcal{A}J(x,t)\right) dt\right] \tag{5.10}$$

In order to fulfil the equality in equation (5.10) we set the arguments of the integral to zero and interchange the maximum operator with the integral and finally arrive at the HJB equation

$$\max_{u\in\mathcal{U}} \left\{L(x,u,t) + \frac{\partial J(x,t)}{\partial t} + \mathcal{A}J(x,t)\right\} = 0, \tag{5.11}$$

and by expanding the differential operator $\mathcal{A}$ we get

$$\begin{aligned} -\frac{\partial J(x,t)}{\partial t} = \max_{u\in\mathcal{U}} \Bigg\{ & L(x,u,t) + \frac{\partial J(x,t)}{\partial x} f(x,u,t) \\ & + \frac{1}{2}\mathrm{tr}\left\{g(x,u,t)g^T(x,u,t)\frac{\partial^2 J(x,t)}{\partial x^2}\right\}\Bigg\}. \end{aligned} \tag{5.12}$$

This is the *Hamilton-Jacobi-Bellman* equation for a stochastic process of (5.1) and the cost functional (5.2). The maximizing $u(t)$ can now be found in terms of x, t, J_x, and J_{xx} and reinserted into (5.12). By solving the resulting PDE for the cost-to-go function $J(x,t)$, the explicit solution for the optimal control $u(t)$ can be found.

5.2.3 Solution Procedure

The following steps lead to the optimal feedback control law as well as to the maximal value function:

1. For a fixed $J(x,t)$, find $u = u(x, J_x, J_{xx}, t)$ such that

   $$L(x,u,t) + J_x(x,t)f(x,u,t) + \frac{1}{2}\mathrm{tr}\{g(x,u,t)g^T(x,u,t)J_{xx}(x,t)\}$$

   is maximal.

2. The function u is put back into the HJB equation for u, and the partial differential equation

   $$J_t(x,t) + L(x,u,t) + J_x(x,t)f(x,u,t) + \frac{1}{2}\mathrm{tr}\{g(x,u,t)g^T(x,u,t)J_{xx}(x,t)\} = 0$$

   is solved with the terminal condition $J(x,T) = K(x)$. The solution $J(x,t)$ is the maximal value function.

3. The function $J(x,t)$ is put back into the equation for u, which is derived in Step 1. This results in the optimal feedback control law: $u^* = u^*(x, J(x,t), t)$.

In the next section, this step by step "cook book" recipe is applied to a linear dynamic system and a quadratic cost functional.

5.2.4 Stochastic LQG Examples with HJB Equation

Below, the stochastic linear quadratic problem is formulated and its solution is given.

Consider the stochastic LQG problem:

$$\begin{aligned} dx(t) &= [A(t)x(t) + B(t)u(t) + b(t)]dt + [C(t)x(t) + \sigma(t)]dW(t) \\ x(0) &= x_0. \\ J(x,t) &= \mathrm{E}\Big\{\frac{1}{2}x^T(T)Fx(T) + g^T x(T) \\ &\quad + \frac{1}{2}\int_0^T \Big(x(t)^T Q(t)x(t) + x(t)^T S(t)u(t) \\ &\qquad + u(t)^T S^T(t)x(t) + u(t)^T R(t)u(t)\Big)dt\Big\} : \min! \end{aligned}$$

where

$$F \geq 0\,,\ R(t) > 0\,,\ Q(t) \geq 0\,,\ \begin{bmatrix} Q(t) & S(t) \\ S^T(t) & R(t) \end{bmatrix} \geq 0\,.$$

We follow the "cook book" recipe of Section 5.2.3.

1. Compute $u(x, J_x, t)$ for a fixed $J(x,t)$:

$$\begin{aligned} J_t + \max_u \Big\{ &\frac{1}{2}(x^T Q x + x^T S u + u^T S^T x + u^T R u) \\ &+ J_x(Ax + Bu + b) + \frac{1}{2}\operatorname{tr}\big\{(Cx+\sigma)(Cx+\sigma)^T J_{xx}\big\}\Big\}. \end{aligned}$$

   The maximization with respect to $u(x, J_x, t)$ yields

$$Ru + S^T x + B^T J_x^T = 0\,,$$

   which gives us

$$u(x, J_x, t) = -R^{-1}\big(S^T x + B^T J_x^T\big)\,.$$

2. Plug the obtained $u(t, J_x, x)$ into the HJB equation and solve it.

$$\begin{aligned} 0 = J_t + \frac{1}{2}\Big\{ &x^T Q x - x^T S R^{-1} S^T x - J_x B R^{-1} B^T J_x^T + x^T (A - BR^{-1}S^T)^T J_x^T \\ &+ J_x(A - BR^{-1}S^T)x + x^T C^T J_{xx} C x + \sigma^T J_{xx}\sigma + x^T C^T J_{xx}\sigma \\ &+ \sigma^T J_{xx} C x + 2b^T J_x^T\Big\}\,. \end{aligned}$$

In order to solve this nonlinear PDE, a quadratic Ansatz is made:

$$J(x,t) = \frac{1}{2}x^T(t)K(t)x(t) + \varphi(t)^T x(t) + \psi(t)\ .$$

The Ansatz is plugged back into the PDE, and the conditions for the solution are derived:

$$\begin{aligned} 0 = \frac{1}{2}x^T\dot{K}x + \dot{\varphi}^T x + \dot{\psi} + \frac{1}{2}x^T\Big(&Q - SR^{-1}S^T - KBR^{-1}B^TK \\ &+(A - BR^{-1}S^T)^TK + K(A - BR^{-1}S^T) + C^TKC\Big)x \\ &+x^T\Big[(A - BR^{-1}S^T - BR^{-1}B^TK)^T\varphi + Kb + C^TK\sigma\Big] \\ &+\frac{1}{2}\sigma^TK\sigma - \frac{1}{2}\varphi^TBR^{-1}B^T\varphi + b^T\varphi\,. \end{aligned}$$

In order to assure that the above equation is satisfied for all $x \in \mathbb{R}^n$, the quadratic term $x^T(...)x$ is set to zero, the linear term $x^T[...]$ is set to zero, and consequently, the constant term is set to zero. The conditions for the quadratic term are

$$\begin{aligned} \dot{K} = &-Q + SR^{-1}S^T + KBR^{-1}B^TK \\ &-(A - BR^{-1}S^T)^TK - K(A - BR^{-1}S^T) - C^TKC \\ K(T) = &\,F\,. \end{aligned}$$

This is a matrix Riccati differential equation. The conditions that the linear and the constant terms vanish are given by

$$\begin{aligned} \dot{\varphi} = &-(A - BR^{-1}S^T - BR^{-1}B^TK)^T\varphi - Kb - C^TK\sigma \\ \varphi(T) = &\,g \end{aligned}$$

and

$$\begin{aligned} \dot{\psi} = &-\frac{1}{2}\sigma^TK\sigma + \frac{1}{2}\varphi^TBR^{-1}B^T\varphi - b^T\varphi \\ \psi(T) = &\,0\,. \end{aligned}$$

3. Since a solution for the HJB PDE is found, the linear state feedback law can now be stated explicitly. We remember from Step 1:

$$u(x, J_x, t) = -R^{-1}\big(S^Tx + B^TJ_x^T\big)\,;$$

and with

$$J_x^T = Kx + \varphi\,,$$

the first order derivative of the quadratic Ansatz $J(x)$ in Step 2, the resulting optimal feedback law is:

$$u^*(x,t) = -R^{-1}(B^TK + S^T)x - R^{-1}B^T\varphi\,.$$

The value of the optimal cost functional can be computed, because the optimal value function has been found:

$$J_0 = J(0, x_0) = \frac{1}{2} x_0^T K(0) x_0 + \varphi(0)^T x_0 + \psi(0)\,.$$

For the LQG Problem, an explicit control law can be found, because of its special structure which has been exploited. In general, finding the solution of a stochastic optimal control law is a tremendous undertaking, since the nonlinear PDE (of Step 2) is hardly ever solvable.

5.2.5 Stochastic Pontryagin's Maximum Principle

Using the knowledge we have from the HJB equation (5.12), we can derive the stochastic Pontryagin's Maximum Principle. Here, a system of forward-backward stochastic differential equations (FBSDE) replaces the HJB partial differential equation. We regard the stochastic optimal control problem as stated in Section 5.2.1.

Before stating the stochastic maximum principle, we need to define the adjoint variables for the stochastic optimal control problem:

$$p(t) = J_x^T(x,t) = \nabla_x J(x,t) \tag{5.13}$$

$$\text{and} \quad p_x(t) = \frac{\partial p}{\partial x} = J_{xx}(x,t) \tag{5.14}$$

and the stochastic version of the Hamiltonian function $H : \mathbb{R}^n \times \mathcal{U} \times \mathbb{R}^n \times \mathbb{R}^{n\times n} \times [0,T]$:

$$\begin{aligned} H(x,u,p,p_x,t) = {} & L(x,u,t) + p^T f(x,u,t) \\ & + \frac{1}{2}\mathrm{tr}\{g(x,u,t) g^T(x,u,t) p_x\}\,. \end{aligned} \tag{5.15}$$

With these definitions, the stochastic version (5.12) of the HJB equation is:

$$-J_t = \max_{u\in\mathcal{U}} H(x,u,p,p_x,t)\,. \tag{5.16}$$

We assume, there is a unique H-minimizing control law $u^* : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^{n\times n} \times [0,T] \to \mathcal{U}$. And in order to avoid cumbersome notation, we introduce the following substitutions:

$$H^*(x,p,p_x,t) = H(x,u^*(x,p,p_x,t),p,p_x,t) \tag{5.17}$$

$$\widetilde{L}(x,p,p_x,t) = L(x,u^*(x,p,p_x,t),t) \tag{5.18}$$

$$\widetilde{f}(x,p,p_x,t) = f(x,u^*(x,p,p_x,t),t) \tag{5.19}$$

$$\widetilde{g}(x,p,p_x,t) = g(x,u^*(x,p,p_x,t),t)\ . \tag{5.20}$$

Thus, we arrive at the following explicit version of the HJB equation:

$$\begin{aligned} -J_t(x,t) &= H^*(x,p,p_x,t) \\ &= \widetilde{L}(x,p,p_x,t) + p^T \widetilde{f}(x,p,p_x,t) \\ &\quad + \frac{1}{2}\,\mathrm{tr}\{\widetilde{g}(x,p,p_x,t)\widetilde{g}^T(x,p,p_x,t)p_x\}\ . \end{aligned} \tag{5.21}$$

In the next step, we write the differential equations for the state x and the costate p. Combining (5.1) and (5.21) we get the following simple expression for dx:

$$\begin{aligned} dx &= f(x,u^*,t) + g(x,u^*,t)dW \\ &= \nabla_p H^*(x,p,p_x,t)dt + \widetilde{g}(x,p,p_x,t)dW\,. \end{aligned} \tag{5.22}$$

Using Itô's lemma on the definition of p in (5.13) yields:

$$dp = J_{xt}^T dt + J_{xx}dx + \frac{1}{2}J_{xxx}(dx)^2 \tag{5.23}$$

$$= \Big[J_{xt}^T + J_{xx}\widetilde{f} + \frac{1}{2}\mathbf{tr}\{\widetilde{g}\,\widetilde{g}^T J_{xxx}\}\Big]dt + J_{xx}\widetilde{g}dW, \tag{5.24}$$

where $\mathbf{tr}\{\widetilde{g}\widetilde{g}^T J_{xxx}\} = [\mathrm{tr}\{(\widetilde{g}\widetilde{g}^T J_{x_1})_{xx}\}, \mathrm{tr}\{(\widetilde{g}\widetilde{g}^T J_{x_2})_{xx}\}, \ldots, \mathrm{tr}\{(\widetilde{g}\widetilde{g}^T J_{x_n})_{xx}\}]^T$.

In order to calculate the term J_{xt}, we differentiate both sides of the HJB equation (5.21) with respect to x and we get

$$\begin{aligned} -J_{xt} &= H_x^* + H_p^* \frac{\partial p}{\partial x} + H_{p_x}^* \frac{\partial p_x}{\partial x} \\ -J_{xt}^T &= \nabla_x H^* + J_{xx}\widetilde{f} + \frac{1}{2}\mathbf{tr}\{\widetilde{g}\widetilde{g}^T J_{xxx}\}\,. \end{aligned} \tag{5.25}$$

Inserting (5.25) into (5.24) leads to

$$dp = -\nabla_x H^* dt + J_{xx}\widetilde{g}dW \tag{5.26}$$

$$= -\nabla_x H^* dt + p_x \widetilde{g}dW\,. \tag{5.27}$$

We can finally write the system of forward-backward stochastic differential equation (FBSDE) for the stochastic maximum principle:

$$\begin{aligned} dx^* &= \nabla_p H^* dt + \widetilde{g}dW \\ dp^* &= -\nabla_x H^* dt + p_x \widetilde{g}dW \\ x^*(0) &= x_0 \\ p^*(T) &= \nabla_x K(x(T))\ . \end{aligned} \tag{5.28}$$

This stochastic two-point-boundary value problem is similar to the deterministic case:

a) Optimal state trajectory obtained from dx^*.
b) Optimal costate trajectory obtained from dp^*.
c) Global static maximization of the Hamiltonian function implied.

5.2.6 Stochastic LQG Example with Maximum Principle

Consider the stochastic LQG problem

$$\begin{aligned} dx &= \big[A(t)x(t) + B(t)u(t)\big]dt + \sigma(t)dW(t) \\ x(0) &= x_0 \\ J(u) &= \mathrm{E}\Big\{\frac{1}{2}x^T(T)Fx(T) \\ &\qquad +\frac{1}{2}\int_0^T \Big(x(t)^T Q(t)x(t) + u(t)^T R(t)u(t)\Big)dt\Big\} \end{aligned}$$

with $F \geq 0$, $Q(t) \geq 0$, $R(t) > 0$. Find the optimal control $u^* : [0,T] \to \mathbb{R}^m$ which minimizes the cost functional $J(u)$.
The Hamiltonian function is:

$$H(t,x,u,p,p_x) = \frac{1}{2}\big(x^T Q x + u^T R u\big) + p^T\big[Ax + Bu\big] + \frac{1}{2}\operatorname{tr}\{\sigma\sigma^T p_x\}\,.$$

Pontryagin's necessary conditions are:

$$\begin{aligned} dx^*(t) &= \nabla_p H_{|*}dt + \sigma dW \\ &= \big[A(t)x^*(t) + B(t)u^*(t)\big]dt + \sigma(t)dW(t) \\ x^*(0) &= x_0 \\ dp^*(t) &= -\nabla_x H_{|*} + p_x^*\sigma dW \\ &= -\big[Q(t)x^*(t) + A^T(t)p^*(t)\big]dt + p_x^*\sigma(t)dW(t) \\ p^*(T) &= Fx^*(T) \\ H(x^*,u^*,p^*,p_x^*,t) &\leq H(x^*,u,p^*,p_x^*,t)\,. \end{aligned}$$

From the last statement we conclude, that

$$\begin{aligned} u^*(t) &= \arg\min_u \frac{1}{2}\Big(u^T(t)R(t)u(t) + u^T(t)B^T(t)p^*(t)\Big) \\ &= -R^{-1}(t)B^T(t)p^*(t)\,. \end{aligned}$$

Plugging this H-minimizing control into the differential equations for x^* and p^* yields the following forward-backward stochastic differential equations:

$$\begin{aligned} dx^*(t) &= \big[A(t)x^*(t) - B(t)R^{-1}(t)B^T(t)p^*(t)\big]dt + \sigma(t)dW(t) \\ x^*(0) &= x_0 \\ dp^*(t) &= -\big[Q(t)x(t) + A^T(t)p^*(t)\big]dt + p_x^*\sigma(t)dW(t)) \\ p^*(T) &= Fx^*(T)\,. \end{aligned}$$

Next, we make a linear Ansatz for $p^*(t) = K(t)x + \varphi(t)$. This makes sense as we can see that $p^*(t)$ is linear in $x^*(t)$. We take the total differential taking stochastic calculus into account:

$$\begin{aligned} dp^* &= \Big[\dot{K}x + \dot{\varphi} + K(Ax - BR^{-1}B^T p^*) + \frac{1}{2}\cdot 0\Big]dt + K\sigma dW \\ &= \Big[\dot{K}x + \dot{\varphi} + K\big(Ax - BR^{-1}B^T(Kx+\varphi)\big)\Big]dt + p_x^*\sigma dW\,. \end{aligned}$$

From the FBSDE system we know

$$dp^* = (-Qx - A^T p^*)dt + p_x^*\sigma dW = \big[-Qx - A^T(Kx+\varphi)\big]dt + p_x^*\sigma dW\,.$$

Combining the two equations and using $p_x = K$ yields:

$$\begin{aligned} &\Big[\dot{K}x + \dot{\varphi} + K\big(Ax - BR^{-1}B^T(Kx+\varphi)\big)\Big]dt + K\sigma dW \\ =\ &\Big[-Qx - A^T(Kx+\varphi)\Big]dt + K\sigma dW \end{aligned}$$

and therefore:

$$\dot{K}x + \dot{\varphi} + K\big(Ax - BR^{-1}B^T(Kx+\varphi)\big) = -Qx - A^T(Kx+\varphi)\,.$$

This finally leads to the following two differential equations for K and φ:

$$\begin{aligned} \dot{K}(t) &= -K(t)A(t) - A^T(t)K(t) + K(t)B(t)R^{-1}(t)B^T(t)K(t) - Q(t) \\ K(T) &= F \\ \dot{\varphi}(t) &= -[A(t) - B(t)R^{-1}(t)B^T(t)K(t)]^T\varphi(t) \\ \varphi(T) &= 0\,. \end{aligned}$$

Note, that this result illustrates two important facts: On one hand we can see that the solution procedure with the FBSDE from the Pontryagins Maximum Principle and the PDE from the HJB equation end up with equivalent results for the optimal control u^*. On the other hand, we see that for our process with linear drift and additive Gaussian noise, the result is the same as in the purely deterministic case. We call this property the *certainty equivalence principle*.

6

Financial Applications

Successful investing is anticipating
the anticipations of others.
John Maynard Keynes (1883-1946)

6.1 Introduction

This section is an introduction to the terms of financial markets. A financial investment, in contrary to a real investment which involves tangible assets (e.g., land, factories), is an allocation of money with contracts in order to make its value increase over time. Conversely, if consumption exceeds the present savings, one has to be willing to pay back more in the future. Therefore, a security is a contract to receive prospective benefits under stated conditions (e.g., stocks, bonds, options).

According to the above statements, an investment can be looked at as a sacrifice of current cash for future cash. The two main attributes that distinguish securities are time and risk, where risk also includes inflation. The interest rate or return is defined as the gain or loss of the investment divided by the initial value of the investment. An investment always contains some sort of risk. Therefore, the higher an investor considers the risk of a security, the higher the rate of return he demands, sometimes referred as risk premium.

The first type of securities considered are *Treasury Bills*. They involve loaning on a short-term basis to the *U.S. Treasury*. Treasury Bills contain almost no risk. The second type of securities are *Bonds*. There are two categories of Bonds: *Government Bonds* and *Corporate Bonds*. Bonds, as Treasury Bills, involve lending money but on a fairly long-term basis. Bonds result in a cash payment each year (the coupon) up to it's expiry, i.e., the maturity date, when the final cash payment (the principal or face value) is made. There is a market for bonds where they can be bought and sold. The price of a bond varies over it's lifetime. The bond rating system is a measure of a company's credit risk. There are several rating agencies such as Moody's, Standard & Poor, and Fitch.

The last type of securities are stocks. Stocks (or shares) represent ownership in a part of a corporation. The board of directors of a corporation decides when to pay cash dividends and how much.

Another important term is short selling. Short selling essentially means selling stocks which are borrowed with the intention to buy them back later. An investor who is short selling (sometimes referred as "shorting") assets believes in declining prices.

Of course it is not possible to cover all terms of investing in this small section. For more details, the reader should refer to [31], [34], and [23].

6.1.1 Continuous Compounding

In the discrete case, the return r of the sample period $[t, t+1]$ on a security with value S is defined by

$$r_t = \frac{S_{t+1} - S_t}{S_t}.$$

Therefore, a return is always calculated with respect to the length of the sample period. In the financial environment, the return is usually calculated on a yearly basis if not stated otherwise (sometimes denoted by p.a.). From the equation above, it follows that $S_{t+1} = S_t(1 + r_t)$. If we assume that the security S has a constant interest rate of r, the value after k periods is calculated as $S_{t+k} = S_t(1+r)^k$. If the sample period is divided into n equal intervals, the return in one such interval would be $\frac{r}{n}$. This yields the following wealth after k periods

$$S_{t+k} = S_t(1 + \frac{r}{n})^{nk}.$$

We now let n tend to infinity and derive the formula for continuous compounding:

$$S_{t+k} = S_t \lim_{n\to\infty}(1 + \frac{r}{n})^{nk} = S_t\ e^{rk}$$

Now, k does no longer need to be an integer and therefore we have

$$S_{t+\triangle t} = S_t\ e^{r\triangle t}.$$

This result will play an important role in the valuation of securities and derivatives in continuous-time.

6.1.2 Net Present Value

This section considers the problem of how to value future cash flows. How should a coupon-paying bond be valued? The concept of the net present value deals with this kind of questions. The *yield to maturity* is used to value bonds.

First of all, we define the present value of a known cashflow (positive or negative) in the future (e.g., bonds). If r is the interest rate then receiving (or paying) S dollars in one year is worth $S\frac{1}{1+r}$ today, in the continuous case it is worth Se^{-r} today. When the cashflow takes place in T years, then its present value is $S\frac{1}{(1+r)^T}$ or Se^{-rT} in the continuous case, Note that r is assumed constant, but the concept can easily be expanded to variable interest rates.

The process of valuing future cashflows as equivalent to present values is called *discounting*.

The net present value (NPV) of an investment (or project) is defined as all discounted future cashflows. If an investor has several possibilities to invest his money, he should invest in the one having the highest net present value.

$$NPV = \sum_{i=1}^{\infty} S_i \Big(\frac{1}{1+r_i}\Big)^i = \sum_{i=1}^{\infty} S_i e^{-r(t_i)t_i} = \int_0^{\infty} S_i(t) e^{-r(t)t} dt$$

When calculating the present value of future cashflows, one needs to make an assumption about the interest rate r. In the case of bonds, we know the present value of the bond since it is sold on the market. Therefore, the only unknown in the equation of the net present value ist the interest rate ($r_i = r = const$). The resulting interest rate is called the *yield to maturity*.

6.1.3 Utility Functions

Suppose an investor has many different investment opportunities that could influence his wealth at the end of the year. Once the investor decides to allocate his capital among the alternatives, his future wealth is governed by the corresponding random cash-flows of the investment opportunities. If all outcomes were certain, it would be easy to rank the options. In an uncertain environment, this decision is not so obvious. Utility functions provide a ranking to judge uncertain situations. For a risk averse investor, a utility function U must fulfill certain properties:

- A utility function must be an increasing continuous function: $U' > 0$
- A utility function must be concave: $U'' < 0$

The first property makes sure that an investor prefers always more wealth to less wealth. The second property captures the principle of risk aversion. Some commonly used utility functions include

1. the exponential function ($a > 0$)
$$U(x) = -e^{-ax}$$
2. the logarithmic function
$$U(x) = \ln(x)$$
3. the power functions ($b < 1$ and $b \neq 0$)
$$U(x) = \frac{1}{b}(x)^b$$
4. the quadratic functions ($x < \frac{a}{2b}$)
$$U(x) = ax - bx^2$$

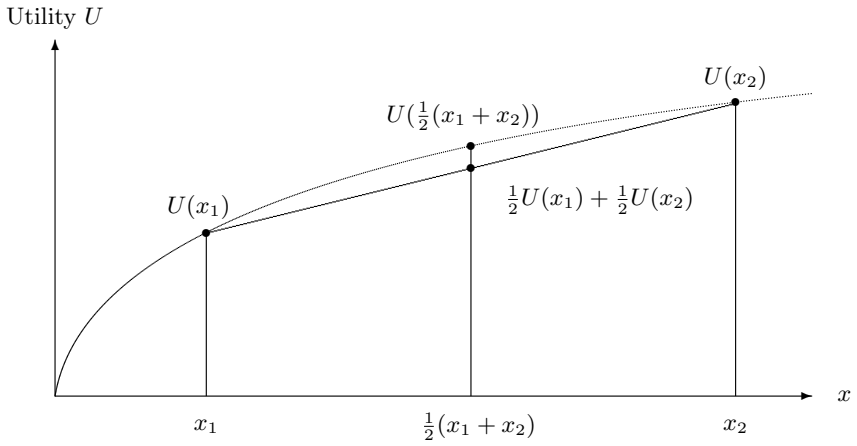

Fig. 6.1. Risk averse utility functions

Any of the utility functions shown above capture the principle of risk aversion. This is accomplished whenever the utility function is concave.

To illustrate this point, we consider two wealth possibilities. The first will either yield with equal probability x_1 or x_2 ($x_2 > x_1$). The second possibility will always yield $\frac{1}{2}(x_1 + x_2)$. To rank both options, the expected utility is computed. The first option yields $\frac{1}{2}[U(x_1) + U(x_2)]$ and the second $U(\frac{1}{2}(x_1 + x_2))$. Since U is concave, the line between $U(x_1)$ and $U(x_2)$ will be below the function of U, and thus the second option has a higher utility, as shown in Figure 6.1.

Both options have the same expected value, but the first option involves more risk, since one might end up with $U(x_1)$. The second option is risk free, since the payoff is deterministic.

Risk aversion coefficients

Since we are not actually interested in the absolute values of utility functions but rather in its shape, Pratt and Arrow have developed measures for risk aversion. These are

- the Arrow-Pratt measure of absolute risk aversion:

$$a(x) = -\frac{U''(x)}{U'(x)}.$$

- the Arrow-Pratt measure of relative risk aversion:

$$b(x) = -x\,\frac{U''(x)}{U'(x)}.$$

6.2 Mean-Variance Portfolio Theory

6.2.1 Introduction

The first attempt to systematically solve the asset allocation problem was the mean-variance approach. An investor usually faces the problem of investing money in n different risky securities. Mean-variance portfolio selection gives a solution to this kind of problem. Mean-variance portfolio selection was introduced by Harry Markowitz in 1952 and is called *modern portfolio theory*. The fundamental assumption of the mean-variance theory is that the returns are normally distributed. Although this is not always the case, it is still a reasonable assumption. The goal of the mean-variance analysis is to find a control vector u, whose elements state how much of the total wealth should be invested in each of the n securities. Therefore, the sum of the elements of u must be one. A portfolio P is fully described by the vector u. Since the returns of the securities are normally distributed, its distributions are fully described by its mean and its variance. The vector u is defined as

$$u = [u_1, u_2, ..., u_n]^T, \quad \sum_{i=1}^{n} u_i = 1.$$

The returns R_i of the n securities are described by their means, denoted by the vector with elements μ_i, and their covariance matrix Σ, respectively. This yields the expected portfolio return

$$\mu_P = \mathrm{E}[R_P] = u^T \mu = \sum_{i=1}^{n} u_i \mu_i$$

and the variance of the portfolio

$$\sigma_P^2 = \mathrm{var}(R_P) = u^T \Sigma u = \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_{ij} u_i u_j \qquad (\sigma_{ii} = \sigma_i^2).$$

With this framework, we can now construct infinitely many possible portfolios. But we want to construct an optimal portfolio in the mean-variance sense. Such a portfolio is also called mean-variance efficient and is defined as follows:

Definition 6.1. *A portfolio u^* is called mean-variance efficient if there exists no portfolio u with $\mu_u \geq \mu_{u^*}$ and $\sigma_u^2 \leq \sigma_{u^*}^2$*

Having introduced the concept of utility, it is natural to maximize the expected end-of-period utility:

$$\begin{aligned} &\max_{u \in \mathbb{R}^n} \quad \mathrm{E}\Big\{U[W_0(1 + R_P)]\Big\} \\ &\quad s.t. \\ &e^T u = 1 \quad \text{with} \quad e = [1, 1, ...1]^T \in \mathbb{R}^n. \end{aligned}$$

In general, there exists no analytical solution for this problem.

6.2.2 The Markowitz Model

The solution provided by Markowitz is very popular and still widely used in practice. Markowitz was the first to include the covariances of the returns in his model. The statement of the optimization of Markowitz is the following:

$$\begin{aligned} \max_{u \in \mathbb{R}^n} \quad & \Big\{ -\frac{1}{2} u^T \Sigma u + \tau \mu^T u \Big\} \\ s.t. \quad & \\ e^T u = 1 \quad & \text{with} \quad e = [1, 1, ...1]^T \in \mathbb{R}^n. \end{aligned}$$

Here, $\tau \geq 0$ measures the investors risk tolerance and is closely related to the Arrow-Pratt measure of relative risk aversion. Every solution of the problem formulation above is a mean-variance efficient portfolio.

To find a solution to this problem we use the method of Lagrange multipliers.

$$\begin{aligned} L(u, \lambda) &= -\frac{1}{2} u^T \Sigma u + \tau \mu^T u + \lambda (e^T u - 1) \\ \frac{\partial L}{\partial u} &= -\Sigma u + \tau \mu + \lambda e = 0 \\ \frac{\partial L}{\partial \lambda} &= e^T u - 1 = 0 \end{aligned}$$

Since Σ is positive-definite, the two conditions are necessary and sufficient conditions for optimality. The solution for λ is

$$\lambda = \frac{1}{e^T \Sigma^{-1} e} - \tau \frac{e^T \Sigma^{-1} \mu}{e^T \Sigma^{-1} e}.$$

This gives the following solution for u:

$$u^* = \Sigma^{-1} \Big(\frac{1}{e^T \Sigma^{-1} e} e + \tau \Big(\mu - \frac{e^T \Sigma^{-1} \mu}{e^T \Sigma^{-1} e} e \Big) \Big).$$

Figure 6.2 shows the result. Every solution for u^* ($\tau > 0$) is mean-variance efficient and lies on the so called efficient set, sometimes also denoted as efficient frontier. The optimal allocation $u^*(\tau = 0)$ has the lowest possible variance and is therefore called the minimum variance point.

By including m additional, arbitrary constraints (e.g., nonnegativity of the elements of u) the optimization looks as follows

$$\begin{aligned} \max_{u \in \mathbb{R}^n} \quad & \Big\{ -\frac{1}{2} u^T \Sigma u + \tau \mu^T u \Big\} \\ s.t. \quad & \\ e^T u = 1 \quad & \text{with} \quad e = [1, 1, ...1]^T \in \mathbb{R}^n \\ Au \leq b \quad & \text{with} \quad A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m. \end{aligned}$$

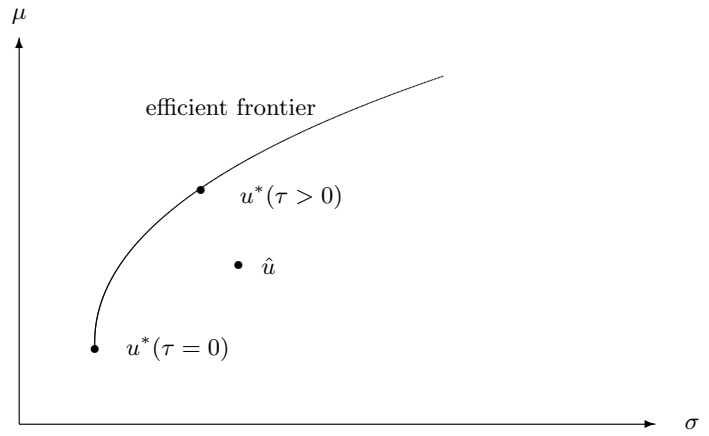

Fig. 6.2. Mean variance plot and efficient frontier

This type of problem is called a quadratic program. There are numerous programs for numerically solve this kind of problems, even for very large dimensions.

Inclusion of a risk-free asset

So far, we have considered only n risky asset with $\sigma > 0$. But what happens if we include a riskless asset with return r and $\sigma = 0$? In this section, we want to answer this question. It is no longer necessary for the u_is to sum to one, we rather invest the difference to one, $(1 - e^T u)$, in the riskless security. The new optimization problem is

$$\max_{u \in \mathbb{R}^n} \Big\{ -\frac{1}{2} u^T \Sigma u + \tau(\mu^T u + (1 - e^T u)r) \Big\}.$$

Since we have no constraints (because they are inherent in the problem) the problem can be solved in a straightforward manner. As Σ is positive-definite, the first order condition is a necessary and a sufficient condition:

$$\tau(\mu - er) - \Sigma u = 0 \quad \Leftrightarrow \quad u = \tau \Sigma^{-1}(\mu - er).$$

Hence the portfolio mean and variance are

$$\begin{aligned} \mu_P &= \tau(\mu - er)^T \Sigma^{-1}(\mu - er) + r \\ \sigma_P^2 &= \tau^2(\mu - er)^T \Sigma^{-1}(\mu - er). \end{aligned}$$

Therefore, the efficient frontier in the $\sigma - \mu$ plane is a straight line, which is shown in Figure 6.3. Intuitively, this makes sense. Every mean-variance

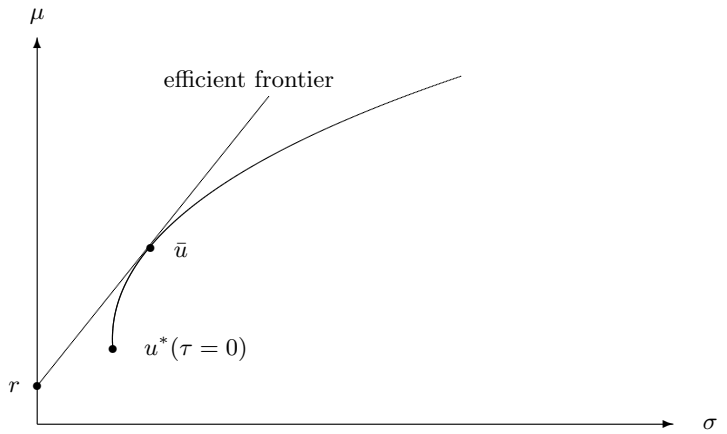

Fig. 6.3. Mean variance plot and efficient frontier

efficient portfolio (without the riskless asset) combined with the riskless asset yields a straight line. Hence, the best one is the tangent on the original efficient set, as shown in Figure 6.3. For a more detailed treatment of the topic of mean-variance, the reader may consult [29] or [23].

6.2.3 The Capital Asset Pricing Model (CAPM)

The capital asset pricing model (CAPM) was developed by Sharpe, Lintner, and Mossin (1964). The assumptions behind the CAPM are pretty daring but still the CAPM is widely used because of its easiness. The main assumptions are:

- Everyone has the same estimations for μ and Σ for the available risky assets and can invest at the same riskfree rate.
- Everyone uses the same mean-variance approach for constructing his portfolio.
- There are no transaction costs or taxes.
- All information about the market is freely and instantly available to all investors.

Under these assumptions, everybody chooses the same tangency portfolio $\bar{u}$ as in Figure 6.3. Therefore, every portfolio is a combination of the portfolio $\bar{u}$ and the riskless asset. In the case of the CAPM, the portfolio $\bar{u}$ is called the market portfolio and it corresponding tangency line is called the capital market line.

All these assumptions require the market to be in equilibrium. The market portfolio consists of all shares available in the market. If a security would not

be in the market portfolio, nobody would buy it and therefore its price would decline until it yields a return that makes it entering the market portfolio again. Therefore it can be stated:

The market portfolio is a portfolio consisting of all securities in which the proportion invested in each security corresponds to its relative market value. The relative market value of a security simply is equal to the aggregate market value[1] *of a security divided by the sum of aggregate market values of all securities, see [34].*

Since every investor holds a portfolio on the capital market line, the expected portfolio return can be calculated from its standard deviation and vice versa.

$$\mu_P = r + \frac{\mu_M - r}{\sigma_M}\sigma_P$$

μ_P : expected return of the portfolio
r : risk-free interest rate
μ_M : expected return of the market
σ_M : standard deviation of the market
σ_P : standard deviation of the portfolio

As we know the relation of the return and the variance of efficient portfolios, we still do not know the relation for individual securities. The relationship is the following:

$$\mu_i - r = \beta_i(\mu_M - r)$$

$$\text{where} \quad \beta_i = \frac{\sigma_{iM}}{\sigma_M^2} = \frac{\rho_{im}\sigma_i\sigma_M}{\sigma_M^2} = \frac{\rho_{im}\sigma_i}{\sigma_M}$$

σ_{iM} : covariance of the i-th asset with the market.

For a proof of this formula see [23].

By combining n arbitrary assets of the market, the resulting portfolio can again be expressed by the CAPM. The beta of the portfolio is

$$\beta_P = \sum_{i=1}^{n} \beta_i u_i \quad \Rightarrow \quad \mu_P = r + \beta_P(\mu_M - r),$$

and therefore, the expected return of the portfolio can easily be calculated.

From the results obtained so far, the return R of a security can be described as follows:

$$R_i = r + \beta_i(\mu_M - r) + \varepsilon_i,$$

where ε_i is an arbitrary random variable with $\mathrm{E}[\varepsilon_i] = 0$ and $\sigma_{\varepsilon_i M} = 0$.

[1] The aggregate market value for a common stock of a company is equal to the current market price of the stock multiplied by the number of shares outstanding.

From this definition, it is obvious that a security has two components of risk:

$$\sigma_i^2 = \beta_i^2 \sigma_M^2 + \sigma_{\varepsilon_i}^2.$$

The first term, $\beta_i^2 \sigma_M^2$, is called systematic risk which is tied to the market. Every security with $\beta \neq 0$ includes this type of risk, hence it cannot be reduced by diversification. The second part, $\sigma_{\varepsilon_i}^2$, is the specific risk for every single security. It is also called non-systematic or idiosyncratic risk. It is unrelated to the market and can therefore be reduced by diversification. The total risk of a portfolio under these assumptions is:

$$\sigma_P^2 = \beta_P^2 \sigma_M^2 + \sigma_{\varepsilon_i P}^2 = \Big(\sum_{i=1}^{n} \beta_i u_i \Big)^2 \sigma_M^2 + \sum_{i=1}^{n} u_i^2 \sigma_{\varepsilon_i}^2.$$

To see the effect of diversification, we assume that the proportion in each of the n securities u_i is $\frac{1}{n}$. Then, we get the following specific variance of the portfolio where $\overline{\sigma_{\varepsilon_i}^2}$ denotes the average variance:

$$\sigma_{\varepsilon_i P}^2 = \sum_{i=1}^{n} \frac{1}{n^2} \sigma_{\varepsilon_i}^2 = \frac{1}{n} \Big(\frac{\sum_{i=1}^{n} \sigma_{\varepsilon_i}^2}{n} \Big) = \frac{1}{n} \sigma_{\varepsilon_i}^2.$$

Therefore, the specific variance will vanish as n tends to infinity.

6.2.4 Arbitrage Pricing Theory (APT)

The Arbitrage Pricing Theory (APT) does not require the same assumptions as the CAPM. The main assumptions of the APT are, that the investors always prefer more return to less return with the same risk, capital markets are perfectly competitive, and that the investment universe is sufficiently large. Before we can state the APT, we need to introduce factor models.

Factor Models

Factor models relate the returns of a security to one or more common factors (e.g., GDP, inflation, interest rates, yield spreads, variances, etc.). Since there is always uncertainty in security returns, there is, for every security, a unique error term, which is uncorrelated to the error terms of the other securities. It is also assumed that the unique error term is uncorrelated with every factor included in the model. Therefore, factor models explain the random return R of a security in the following manner:

$$
\begin{aligned}
R_i &= a_i + \sum_{j=1}^{m} b_{ij} F_j + \varepsilon_i \\
R_i &: \text{ the return of the i-th asset} \\
a_i, b_{ij} &: \text{ constants} \\
F_j &: \text{ a factor} \\
\varepsilon_i &: \text{ error term with } \mathrm{E}[\varepsilon_i] = 0.
\end{aligned}
$$

The a_is are sometimes denoted by *intercepts* and the b_{ij}s are denoted by *factor loadings*.

Remember that the CAPM is a one-factor model, where the only factor is the return of the market. The only difference is that the CAPM is an equilibrium model whereas factor models usually are not. This is because the intercept in the CAPM is fixed, i.e., the riskfree rate, but in a general factor model, the intercepts may vary for different securities. The weights of the different factors determines the expected return, variance, and covariances of every security.

As in the CAPM, the diversification effect also reduces the specific risk of a portfolio. This is illustrated as follows. Suppose a portfolio of n securities is constructed with proportion u_i of total wealth invested in the i-th security. It is also assumed that the portfolio is well diversified, i.e., there is no dominating u_i. This yields the random rate of return R_P of the portfolio:

$$
\begin{aligned}
R_P &= \sum_{i=1}^{n} u_i a_i + \sum_{j=1}^{m} b_j F_j + \varepsilon \\
b_j &= \sum_{i=1}^{n} u_i b_{ij} \\
\sigma_\varepsilon^2 &= \sum_{i=1}^{n} u_i^2 \sigma_{\varepsilon_i}^2.
\end{aligned}
$$

We define $\bar{\sigma}_\varepsilon^2 = \max(\sigma_{\varepsilon_i}^2)$, $\bar{u} = \max(u_i)$. Since we demand that there is no dominating u_i we know that $\sum_{i=1}^{n} \bar{u} \approx 1$ and therefore $\bar{u} \approx \frac{1}{n}$ and get for the specific portfolio variance

$$
\sigma_\varepsilon^2 \leq \sum_{i=1}^{n} \frac{1}{n^2} \bar{\sigma}_\varepsilon^2 = \frac{1}{n} \bar{\sigma}_\varepsilon^2.
$$

Therefore, the systematic risk tends to zero for a well diversified portfolio as n tends to infinity. This argument is inherent in the APT.

Statement of the APT

The APT theory was developed by Ross in 1976. APT states that prices are generated by a factor model but does not state which factors. APT extends

the factor model approach such that the market is in equilibrium again. In order for the market to be in equilibrium, there must not be any arbitrage opportunities. This is the main argument to construct the theory.

As shown in the previous chapter, we can omit the specific risk in a well diversified portfolio. As a consequence, the uncertainty stems only from the uncertainty of the factors. Therefore, the return can be expressed as

$$R_i = a_i + \sum_{j=1}^{m} b_{ij} F_{ij} \quad , \quad \mathrm{E}[F_{ij}] = 0.$$

Suppose, we hold a portfolio u. We now state that there is no way to change this portfolio by a so called *arbitrage portfolio* $\triangle u$ which yields a higher return but bears the same risk as the original portfolio. The arbitrage portfolio $\triangle u$ has the following characteristics:

$$\begin{aligned}
\sum_{i=1}^{n} \triangle u_i = 0 \qquad & \triangle u \text{ is self financing.} \\
\sum_{i=1}^{n} \triangle u_i b_{ij} = 0 \qquad & \triangle u \text{ contains no factor risk } j = 1...m. \\
\sum_{i=1}^{n} \triangle u_i a_i = 0 \qquad & \triangle u \text{ must not have a return.}
\end{aligned}$$

We therefore have the following system of $m + 2$ linear equations:

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ a_1 & a_2 & \cdots & a_n \\ b_{11} & b_{21} & \cdots & b_{n1} \\ \vdots & \vdots & \ddots & \vdots \\ b_{1m} & b_{2m} & \cdots & b_{nm} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Obviously the matrix of coefficients must be singular. Therefore we can express any row of the system as a linear combination of the remaining rows, i.e., there exists $\lambda_0, \lambda_1, ..., \lambda_m$ such that

$$a_i = \lambda_0 + \sum_{j=1}^{m} b_{ij} \lambda_j.$$

Since $\mathrm{E}[F_{ij}] = 0$ we know have the following expected return,

$$\mathrm{E}[R_i] = \lambda_0 + \sum_{j=1}^{m} b_{ij} \lambda_j.$$

If there is a riskless asset with return r, we know that $\lambda_0 = r$. The APT simply states that given all b_{ij}, there is no way to chose the a_is, they have to be calculated with the equations above. The λ_js are called factor risk premium, that is the expected excess return on a portfolio that has unit sensitivity to the corresponding factor.

6.3 Continuous-Time Finance

6.3.1 Introduction

Continuous-time finance was developed in the late 1960. Paul Samuelson and Robert Merton were among the main contributors. The breakthrough was in 1973 when Fisher Black and Myron Scholes presented their formula for pricing options. There is vast research today in the area of continuous-time finance. In the context of continuous-time finance, stochastic optimal control plays a very important role.

6.3.2 The Dynamics of Asset Prices

This section introduces models to describe asset price dynamics. The model usually consists of some type of stochastic differential equation. The following assumptions are necessary to develop the models:

- Trading is continuous.
- There are no transaction costs, fees, or taxes.
- The investor cannot influence prices by buying or selling an asset.
- The investor can do short selling.
- The investor can borrow and invest money at the same deterministic interest rate.

To develop the model, we consider a market in which n risky assets exist. The asset price processes $S_i(t)$ of the risk-bearing investments satisfy the stochastic differential equation

$$\begin{aligned} \frac{dS_i(t)}{S_i(t)} &= \mu_i(x(t),t)\ dt + \sigma_i(x(t),t)\ dW_S(t) \\ S_i(0) &> 0\,, \end{aligned}$$

$\mu_i \in \mathbb{R}$ is the relative change in price of S_i and $\sigma_i\sigma_i^T$ is the covariance per unit time ($\sigma_i \in \mathbb{R}^{1\times n}$ is the i-th row of the matrix $\sigma(x(t),t) \in \mathbb{R}^{n\times n}$). The n-dimensional Brownian motion dW_S is defined on a fixed, filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbf{P})$. In general, μ_i and σ_i can be dependent on some arbitrary factors $x(t)$, which may be described by SDEs themselves and the time t. Therefore, μ_i and σ_i may be stochastic processes. The type of stochastic differential equation is a generalized geometric Brownian motion with possibly stochastic drift and diffusion. This model reflects important characteristics of real stock or bond prices (e.g., no negative prices are observed). In the case when $\mu_i(x(t),t)$ and $\sigma_i(x(t),t)$ are constant, the price process is the famous geometric Brownian motion. By adding a further, riskless asset (bank account) with a deterministic interest rate r, with the following dynamics

$$\begin{aligned} dB(t) &= B(t)r(t,x(t))\ dt \\ B(0) &> 0\,. \end{aligned}$$

The process $x(t)$ allows us to model variables which describe external influence factors of either macroeconomic, industry specific, or company specific nature. Examples of these factors are

Macroeconomic factors	Industry specific	Company specific
GDP growth	sector growth	dividends
long term interest rate	industry rate of returns	earnings
inflation	industry levarage	cash-flow

By carefully selecting external variables with some predictive capacity, we can model the time-varying features of the asset price dynamics. A further advantage is that these variables are measurable and are not contained invisibly in the price dynamics. The framework is very broad and thus allows us to model different types of securities, such as stocks, bonds, options, etc. The setup of equations imposes a structure to model asset price dynamics. The price processes are divided into a function for instantaneous expected returns $\mu_i(x(t), t)$ and a function for instantaneous volatility $\sigma_i(x(t), t)$. Both are functions of the external economic variables $x(t)$ and thereby we are able to model the influences of the economic factors on the instantaneous expected returns and volatility of each individual security. The relationships between the external economic variables and the assets price dynamics as well as the dynamics of the external variables can be derived by a theoretical model or by empirical methods applied to financial data. For example, a popular interest rate model could be taken, which theoretically proposes interest rate dynamics and the relationship between the interest rates and the price dynamics of bonds. Alternatively, empirical time series modelling could be used to establish the relationship between asset prices and external economic factors.

6.3.3 Wealth Dynamics and Self-Financing Portfolios

The derivation of wealth dynamics of a portfolio was first developed in Merton, see [25]. In this context, we will use a slightly different approach to develop the wealth dynamics. The only way an investor can change the wealth dynamics is by buying or selling shares. Once the portfolio is constructed and no shares are either bought or sold, the wealth changes according to the changes of the share prices. Therefore, we have the fundamental relationship which holds in the discrete time and in the continuous time case

$$X(t) = \sum_{i=1}^{n} N_i(t) S_i(t)$$

$X(t)$: wealth of the investor at time t

$N_i(t)$: number of shares of the i-th security in the portfolio

$S_i(t)$: share price of the i-th security

n : the number of existing shares on the exchange .

Consider the situation where the investor can only change his portfolio in discrete time and during the sampling period Δt he cannot change the portfolio. If we also allow the investor to take out money of the portfolio we get the following budget equation

$$X(t+\Delta t) - X(t) = \sum_{i=1}^{n} N_i(t)[S_i(t+\Delta t) - S_i(t)] - C(t+\Delta t)\Delta t$$

$C(t) > 0$: consumption per unit time in the interval $[t,\ t+\Delta t]$.

Intuitively, the above equation states that stochastic changes in wealth may only occur from changes in the share prices. Letting $\Delta t \to 0$ and the budget equation yield

$$dX = \sum_{i=1}^{n} N_i(t) dS_i - C(t)dt.$$

We have already modelled the dynamics of the share prices (omitting the time arguments)

$$dS_i = S_i \mu_i \ dt + S_i \sigma_i \ dW_S.$$

Putting this into the wealth equation we get

$$dX = \Big(\sum_{i=1}^{n} N_i S_i \mu_i - C \Big) \ dt + \sum_{i=1}^{n} N_i S_i \sigma_i \ dW_S.$$

As in the mean-variance portfolio theory we define the elements u_i of the control vector u as the fraction of wealth invested in the i-th security. Therefore, we have $u_i = \frac{N_i S_i}{X}$ whereas u_0 denotes the proportion of wealth invested at the risk-free rate r ($\sigma_0 = 0$). This yields the wealth equation

$$dX = \Big(\sum_{i=0}^{n} u_i X \mu_i - C \Big) \ dt + \sum_{i=1}^{n} u_i X \sigma_i \ dW_S.$$

By definition, we have $\sum_{i=0}^{n} u_i = 1$, therefore $u_0 = 1 - \sum_{i=1}^{n} u_i$, so the wealth equation becomes

$$dX = \Big(Xr + X \sum_{i=1}^{n} u_i(\mu_i - r) - C \Big) \ dt + X \sum_{i=1}^{n} u_i \sigma_i \ dW_S.$$

In vector notation with $e = (1, ..., 1)^T \in \mathbb{R}^n$ we get

$$dX = [X(u^T(\mu - e\, r) + r) - C] \ dt + X u^T \sigma dW_S. \tag{6.1}$$

By including the arguments, the equation becomes much harder to read:

$$\begin{aligned} dX(t) = {} & u(t)^T(\mu(x(t),t) - e\, r(t,x(t)))X(t)\, dt + r(t,x(t))X(t)\, dt \\ & + X(t)u(t)^T \sigma(x(t),t) dW_S(t) - C(t)dt. \end{aligned}$$

where $u(t) = (u_1(t), \ldots, u_n(t))^T \in \mathbb{R}^n$, initial condition $X(0) = X_0$, $dW_S(t) \in \mathbb{R}^n$ and

$$\mu(x(t), t) = (\mu_1(x(t), t), \ldots, \mu_n(x(t), t))^T \in \mathbb{R}^n$$
$$\sigma(x(t), t) = (\sigma_1(x(t), t), \ldots, \sigma_n(x(t), t))^T \in \mathbb{R}^{n \times n}$$

Before we can conclude the section there is still one point which needs clarification. If the original wealth equation, $X(t) = \sum_{i=1}^{n} N_i(t)S_i(t)$, is differentiated according to Itô we get:

$$dX(t) = \sum_{i=1}^{n} N_i(t)dS_i(t) + \sum_{i=1}^{n} dN_i(t)S_i(t) + \sum_{i=1}^{n} dN_i(t)dS_i(t).$$

The first term on the right hand side reflects capital gains, as in our discrete model. Therefore, the other two terms on the right hand side must be the consumption of the investor, which we will show now. Again, we consider the budget equation, but set $W(t)$ to $\sum_{i=1}^{n} N_i(t)S_i(t)$

$$\sum_{i=1}^{n} N_i(t + \Delta t)S_i(t + \Delta t) - \sum_{i=1}^{n} N_i(t)S_i(t)$$
$$= \sum_{i=1}^{n} N_i(t)[S_i(t + \Delta t) - S_i(t)] - C(t + \Delta t)\Delta t.$$

Be rearranging the terms we get:

$$-C(t + \Delta t)\Delta t = \sum_{i=1}^{n} [N_i(t + \Delta t) - N_i(t)]S_i(t + \Delta t)$$
$$= \sum_{i=1}^{n} [N_i(t + \Delta t) - N_i(t)][S_i(t + \Delta t) - S_i(t)]$$
$$+ \sum_{i=1}^{n} [N_i(t + \Delta t) - N_i(t)]S_i(t).$$

For $\Delta t \to 0$ we arrive at

$$-C(t)dt = \sum_{i=1}^{n} dN_i(t)dS_i(t) + \sum_{i=1}^{n} dN_i(t)S_i(t).$$

In [25], Merton introduces non-capital gains (wage) addition to the consumption, denoted by dy. In this context, we forego non-capital gains and therefore set dy to zero. This concludes the derivation of the wealth equation (6.1).

6.3.4 Portfolio Models and Stochastic Optimal Control

Example 1: Merton's Problem

We consider a market in which $n + 1$ assets are traded continuously. One of the assets is a bond whose price process $B(t)$ is given by the deterministic

differential equation

$$\begin{aligned} dB &= r(t)B(t)\ dt \\ B(0) &= b_0 > 0\,, \end{aligned}$$

where $r(t) > 0$ is called the interest rate of the bond. As time passes, $B(t)$ will grow steadily and the bond is therefore called a riskless asset.

The other n assets are called stocks. Their price processes $S_1(t), \ldots, S_n(t)$ satisfy the stochastic differential equations

$$\begin{aligned} dS_i(t) &= \mu_i(t)S_i(t)\ dt + \sigma_i(t)^T S_i(t)\ dW(t) \\ S_i(0) &= s_i > 0 \end{aligned}$$

where $\mu \in \mathbb{R}^n$ is called the appreciation rate (drift) and $\sigma_i \in \mathbb{R}^m$ ($\sigma \in \mathbb{R}^{n\times m}$) the volatilities (diffusion) which reflect the fluctuations of the stock prices. The m-dimensional Brownian motion $dW(t)$ is defined on a fixed, filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbf{P})$.

Now, we can apply the wealth equation (6.1):

$$dX = [X(u^T(\mu - e\,r) + r) - C]\ dt + Xu^T\sigma dW,$$

where $u(t) \in \mathbb{R}^n$ is the usual control vector. The total wealth at time $t \geq 0$ is given by $X(t)$.

As a part of the model, we include the investor's objective function. The expected discounted objective (using a weighting factor π) is the following:

$$J(u) = E\left\{ \int_0^T \frac{1}{\gamma} e^{-\rho t} C(t)^\gamma dt + \pi \frac{1}{\gamma} X(T)^\gamma \right\}.$$

Solution of Example 1

We first state the optimization problem:

$$\begin{aligned} \max_u \quad & E\left\{ \int_0^T \frac{1}{\gamma} e^{-\rho t} C(t)^\gamma dt + \pi \frac{1}{\gamma} X(T)^\gamma \right\} \\ s.t. \quad & \\ & dX = [X(u^T(\mu - e\,r) + r) - C]\ dt + Xu^T\sigma dW \\ & X(0) = x_0 > 0\,. \end{aligned}$$

We define the optimal cost-to-go function by

$$J(X(t), t) = \max_{C,u(t)} E\left\{ \int_0^T \frac{1}{\gamma} e^{-\rho t} C(t)^\gamma dt + \pi \frac{1}{\gamma} X(T)^\gamma \right\}.$$

This problem leads to the following HJB equation ($\frac{\partial J}{\partial t} \equiv J_t$, $\frac{\partial J}{\partial X} \equiv J_x$, and $\frac{\partial^2 J}{\partial X^2} \equiv J_{xx}$)

$$J_t + \max_{C(t),u(t)} \left[e^{-\rho t} C^\gamma + \big(X(u^T(\mu - e\, r) + r) - C\big) J_x + \frac{1}{2} X^2 J_{xx} u^T \Sigma u \right] = 0$$

with $\Sigma = \Sigma(t) = \sigma(t)\sigma(t)^T$. To derive the optimal C and u, we first do the maximization as demanded by the equation above (assuming the partial derivatives as constant)

$$u^*(t) = -\frac{J_x}{X J_{xx}} \Sigma^{-1}(\mu - er)$$
$$C^*(t) = \Big(\frac{1}{\gamma} e^{\rho t} J_x\Big)^{\frac{1}{\gamma - 1}}.$$

We now put these value back into the HJB equation

$$J_t + e^{\frac{\rho t}{\gamma - 1}} J_x^{\frac{\gamma}{\gamma - 1}} \Big(\gamma^{\frac{-\gamma}{\gamma - 1}} - \gamma^{\frac{-1}{\gamma - 1}}\Big) - \frac{1}{2} \frac{J_x^2}{J_{xx}} (\mu - er)^T \Sigma^{-1} (\mu - er) + X r J_x = 0$$

We use the following Ansatz

$$J(X, t) = X^\gamma e^{-\rho t} h(t)$$

and get

$$\frac{\partial J}{\partial t} = e^{-\rho t} X^\gamma \big(h'(t) - \rho h(t)\big)$$
$$\frac{\partial J}{\partial X} = e^{-\rho t} h(t) \gamma X^{\gamma - 1}$$
$$\frac{\partial^2 J}{\partial X^2} = e^{-\rho t} h(t) \gamma(\gamma - 1) X^{\gamma - 2}.$$

Inserting these partial derivatives into equations for C^* and u^* yields the optimal policies for consumption $C^*(t)$ and investment strategy $u^*(t)$:

$$u^*(t) = -\frac{1}{\gamma - 1} \Sigma^{-1}(b - er) = \frac{1}{1 - \gamma} \Sigma^{-1}(\mu - er)$$
$$C^*(t) = \Big(h(t) \gamma X^{\gamma - 1}\Big)^{\frac{1}{\gamma - 1}} = h(t)^{\frac{1}{\gamma - 1}} X.$$

Plugging the optimal policies and the partial derivatives from above into the HJB equation results in

$$e^{-\rho t} X^\gamma \left(h'(t) + h(t) \Big(r\gamma - \rho + \frac{(\mu - er)^T \Sigma^{-1} (\mu - er)}{2(1 - \gamma)} \Big) - (1 - \gamma) h(t)^{\frac{\gamma}{\gamma - 1}} \right) = 0.$$

Thus, to specify $h(t)$ and find an explicit solution to the optimal control problem, the ordinary differential equation

$$\dot{h}(t) + A h(t) - (1 - \gamma) h(t)^{\frac{\gamma}{\gamma - 1}} = 0$$
$$h(T) = \pi e^{\rho T},$$

with $A = r\gamma - \rho + \frac{(\mu - er)^T \Sigma^{-1} (\mu - er)}{2(1 - \gamma)}$ remains to be solved.

Example 2: Stochastic Returns

The second example is an extension of the geometric Brownian motion model. The assumption that the parameters μ and σ are constant values are dropped and instead we model the drift term as a stochastic process. The drift is modelled as a mean-reverting stochastic process given by

$$\begin{aligned} d\mu(t) &= \kappa(\theta - \mu(t))\ dt + \sigma_\mu\ dW_\mu \\ \mu(0) &= \mu_0 > 0\,. \end{aligned}$$

The stock market portfolio process is modelled by a geometric Brownian motion, as given by

$$\begin{aligned} dP(t) &= \mu(t)P(t)\ dt + \sigma_P P(t)\ dW_P \\ P(0) &= p_0 > 0\,. \end{aligned}$$

The correlation between Z_μ and Z_P is denoted by ρ. We assume that an investor has the choice to invest in either the stock market or in a short term bond (or money market account), described by

$$\begin{aligned} dB &= rB(t)\ dt \\ B(0) &= b_0 > 0\,, \end{aligned}$$

where r denotes the short term risk-free interest rate and $B(t)$ the price of the continuously compounding bond. Let $u(t)$ the fraction of the wealth invested in the stock market and $1 - u(t)$ the fraction of wealth invested in the short term bond. Again, we apply the wealth equation (6.1) and set $C(t) \equiv 0$:

$$\begin{aligned} dX(t) &= X(t)[u(t)(\mu(t) - r) + r]\ dt + X(t)u(t)\sigma_P\ dW_P \\ X(0) &= x_0 > 0\,. \end{aligned}$$

The wealth dynamics and the dynamics of $\mu(t)$ govern the portfolio of the investor. The optimization problem is to choose $u(t)$ in such way that a risk adjusted performance measure is maximized. This example features only one stochastic asset, but this framework can easily be extended to many assets including many drift term processes.

The investors objective function is assumed to be the following

$$J(u) = E\left\{\frac{1}{\gamma}X(T)^\gamma\right\}.$$

Solution of Example 2

We first state the optimal control problem

$$\begin{aligned}
\max_{u(\cdot)} \quad & E\left\{\frac{1}{\gamma}X(T)^{\gamma}\right\} \\
s.t. & \\
dX(t) &= X(t)[u(t)(\mu(t)-r)+r]\ dt + X(t)u(t)\sigma_P\ dW_P \\
d\mu(t) &= \kappa(\theta-\mu(t))dt + \sigma_\mu dW_\mu \\
dW_P dW_\mu &= \rho \\
X(0) &= x_0 > 0 \\
\mu(0) &= \mu_0 > 0\,,
\end{aligned}$$

where we assume that the investor does not consume any of his wealth until the end of the time horizon. We define the optimal cost-to-go function by

$$J(t,X,\mu) = \max_{u(\cdot)} E\left\{\frac{1}{\gamma}X(T)^{\gamma}\right\}.$$

Before stating the HJB equation in the usual manner, we reformulate the state equations as follows:

$$\begin{aligned}
dX(t) &= X(t)[u(t)(\mu(t)-r)+r]\ dt + X(t)u(t)\sigma_P\ dW_P \\
d\mu(t) &= \kappa(\theta-\mu(t))dt + \sigma_\mu(\rho dW_P + \sqrt{1-\rho^2}dW_\mu).
\end{aligned}$$

These state dynamics leads us to the following HJB equation:

$$\begin{aligned}
J_t + \max_{u(t)} \Big[& J_x X(t)[(\mu-r)u(t)+r] + J_\mu\kappa(\theta-\mu(t)) \\
& + \frac{1}{2}J_{xx}X^2(t)u^2(t)\sigma_P^2 + J_{x\mu}X(t)u(t)\rho\sigma_P\sigma_\mu + \frac{1}{2}J_{\mu\mu}\sigma_\mu^2\Big] = 0 \\
& J(T) = \frac{1}{\gamma}X(T)^{\gamma}
\end{aligned}$$

The optimal $u^*(t)$ has to satisfy the first order conditions:

$$u^*(t) = -\frac{J_x(\mu(t)-r) + J_{x\mu}\rho\sigma_P\sigma_\mu}{J_{xx}X(t)\sigma_P^2}.$$

The optimal $u^*(t)$ is plugged back into the HJB equation and we obtain:

$$\begin{aligned}
& J_t + J_x X(t)r + J_\mu\,\kappa(\theta-\mu(t)) + \frac{1}{2}J_{\mu\mu}\,\sigma_\mu^2 \\
& -\frac{1}{2}\frac{J_x^2(\mu(t)-r)^2}{J_{xx}\sigma_P^2} - \frac{J_{\mu x}J_x(\mu(t)-r)\rho\sigma_\mu}{J_{xx}\sigma_P} - \frac{1}{2}\frac{J_{\mu x}^2\rho^2\sigma_\mu^2}{J_{xx}} = 0.
\end{aligned}$$

We now guess that the value function J has a form

$$J(t,X,\mu) = \frac{1}{\gamma}(X(t)e^{r(T-t)})^{\gamma}\ e^{a(t)+b(t)\mu(t)+c(t)\mu^2(t)}$$

In order to satisfy the terminal condition $J(T, X, \mu) = \frac{1}{\gamma} X^{\gamma}(T)$, the following conditions are imposed

$$a(T) = b(T) = c(T) = 0.$$

The partial derivatives of the value function J are computed and put back into HJB equation. The task of solving a nonlinear PDE is now turned into a task of solving three coupled ODEs for $a(t)$, $b(t)$, and $c(t)$. The following ODEs are obtained:

$$\frac{1}{\gamma}\dot{c}(t) - \frac{2}{\gamma}\kappa c(t) + \frac{2}{\gamma}c^2(t) - \frac{1}{2(\gamma-1)\sigma_P^2} - \frac{2\,\rho\,\sigma_\mu\,c(t)}{(\gamma-1)\sigma_P} - \frac{2\,c^2(t)\,\rho^2\sigma_\mu^2}{(\gamma-1)} = 0$$

with terminal condition $c(T) = 0$,

$$\frac{1}{\gamma}\dot{b}(t) + \frac{\kappa}{\gamma}(2\,\theta\,c(t) - b(t)) + \frac{b(t)\,c(t)}{\gamma} + \frac{2r}{(\gamma-1)\,\sigma_P^2} - \frac{\rho\,\sigma_\mu(a(t) - r\,c(t)}{(\gamma-1)\,\sigma_P} - \frac{2\,\rho^2\,\sigma_\mu^2\,b(t)\,c(t)}{(\gamma-1)} = 0$$

with terminal condition $b(T) = 0$,

$$\frac{1}{\gamma}\dot{a}(t) + \frac{\kappa\,\theta}{\gamma}b(t) + \frac{b^2(t) + 2c^2(t)}{2\gamma} - \frac{r^2}{2(\gamma-1)\sigma^2} - \frac{\rho\,\sigma_\mu\,r\,a(t)}{(\gamma-1)\,\sigma_P^2} - \frac{\rho^2\sigma_\mu^2\,b^2(t)}{2(\gamma-1)} = 0$$

with terminal condition $a(T) = 0$. The optimal portfolio control law is given by

$$u^*(t, X, \mu) = \frac{(\mu(t) - r) + \rho\,\sigma_P\,\sigma_\mu[b(t) + 2\,c(t)\,\mu(t)]}{(1-\gamma)\,\sigma_P^2}.$$

The three ODEs can be solved offline before the controller is used to make the investment decisions.

Example 3: Stochastic Interest Rates

The third model assumes that the interest rate is a stochastic process rather than a given constant. Instead of investing in equities (stocks), we assume the investor invests in long term bonds. The time-varying short-term interest rate is given by

$$\begin{aligned} dr(t) &= \kappa(\theta - r(t))dt + \sigma_r dW \\ r(0) &= r_0 > 0\,, \end{aligned}$$

where $r(t)$ is the short-term interest rate, θ the long run average, $\kappa > 0$ the rate of reversion, r_0 the initial condition, and σ the variance parameter. This model is known as the Vasicek model. Let $B(r(t), \tau)$ denote the price

of a zero-coupon bond with $\tau = T - t$ periods to maturity. The dynamics of bond returns can be derived using the no-arbitrage argument together with a constant price of risk

$$\begin{aligned}\frac{dB(r(t),\tau)}{B(r(t),\tau)} &= [r(t) + \lambda a(\tau)]dt - a(\tau)\sigma_r dW \\ B(0) &= b_0 > 0\,,\end{aligned}$$

where dW denotes the Brownian motion, λ the price of interest rate risk, and $a(\tau) = \frac{1}{\kappa}(1 - e^{-\kappa\tau})$. The second investment possibility is a short-term money market account M which is described by

$$\begin{aligned}dM(t) &= r(t)M(t)dt \\ M(0) &= m_0 > 0\,.\end{aligned}$$

The investor is assumed to invest only in the long-term bond or short-term money market account which pays the short-term interest rate $r(t)$. Let $u(t)$ denote the fraction of the wealth invested in the long-term bond and $1 - u(t)$ the fraction of his wealth put into the money account. The wealth dynamics are given by

$$\begin{aligned}dX &= X(t)u(t)\frac{dB(t)}{B(t)} + (1 - u(t))X(t)\frac{dM(t)}{M(t)} \\ &= u(t)X(t)\Big([r(t) + \lambda a(\tau)]dt - a(\tau)\sigma_r dW\Big) + (1 - u(t))r(t)X(t)dt \\ &= [(u(t)\lambda a(\tau) + r(t))X(t) - C(t)]dt - u(t)X(t)\lambda a(\tau)\sigma_r dW\end{aligned}$$

where $X(t)$ denotes the investor's wealth and $C(t)$ denotes the consumption out of the portfolio.

Solution of Example 3

For this model we assume that the investor wants to maximize his consumption from the portfolio. Instead of maximizing his wealth for a given horizon, we assume that the investor wants to have maximum discounted risk-adjusted cashflow of his investments. The optimal control problem is therefore given by

$$\begin{aligned}\max_{C(\cdot)\geq 0,\, u(\cdot)} \quad & E\left\{\int_0^\infty e^{-\beta t}\frac{C^{1-\gamma}}{1-\gamma}\,dt\right\} \\ s.t.& \\ dX &= [(u(t)\lambda a(\tau) + r(t))W(t) - C(t)]dt - u(t)X(t)a(\tau)\sigma_r dW_P \\ dr(t) &= \kappa(\theta - r(t))dt + \sigma_r dW_r \\ r(0) &= r_0 > 0 \\ X(0) &= x_0 > 0\,,\end{aligned}$$

where we assume that the consumption is positive, $0 < \gamma < 1$, β is the discount rate, and the optimal portfolio selection is unrestricted. The cost-to-go function is defined as

$$J(X,r,t) = \max_{C(\cdot)\geq 0,\, u(\cdot)} E\left\{ \int_0^\infty e^{-\beta t} \frac{C^{1-\gamma}}{1-\gamma}\, dt \right\}$$

For the optimal portfolio and consumption strategy for this model we obtain the following HJB equation:

$$J_t + \sup_{u(t),C(t)} \Big[e^{-\beta t} \frac{C^{1-\gamma}}{1-\gamma} + J_x(u(t)\lambda a(\tau)X(t) + rX(t) - C(t)) + J_r\kappa(\theta - r(t)) \\ + \frac{1}{2} J_{xx} u^2(t) X^2(t) a^2(\tau)\sigma_r^2 - J_{xr} u(t) X(t) a(\tau)\sigma_r^2 + \frac{1}{2} J_{rr}\sigma_r \Big] = 0$$

From the HJB equation, the first-order conditions for optimal consumption and portfolio choice are:

$$C(t) = (e^{\beta t} J_x)^{-\frac{1}{\gamma}}$$
$$u(t) = -\frac{J_x}{J_x X} \frac{\lambda}{a(\tau)\sigma_r^2} + \frac{J_{xr}}{J_{xx} X} \frac{1}{a(\tau)}.$$

Substituting these conditions into the HJB equation gives a second-order nonlinear PDE for the value function J. To solve this equation, we guess that the solution takes the form

$$J(X,r,t) = e^{-\beta t} H(r(t))^\gamma \frac{X(t)^{1-\gamma}}{1-\gamma},$$

where $H(r(t))$ is a function of the short-term interest rate. Using the guess we arrive at an ODE of the form

$$0 = \frac{\gamma}{(1-\gamma)H} + \frac{\lambda^2}{2\gamma\sigma_r^2} - \frac{\beta}{1-\gamma} + r + \Big(\frac{\gamma\kappa(\theta - r)}{1-\gamma} - \lambda\Big)\frac{H'}{H} + \frac{\gamma\sigma_r^2}{2(1-\gamma)}\frac{H''}{H},$$

where H' denotes the derivative with respected to r. The ODE can analytically be solved, but it involves complicated expressions involving gamma functions, and it can be shown that $H(r) \geq 0$. The optimal portfolio control law and the optimal consumption laws are given by

$$u(t) = \frac{\lambda}{a(\tau)\sigma_r^2\gamma} - \frac{1}{H(r(t))a(\tau)}$$
$$C(t) = \frac{X(t)}{H(r(t))}\,.$$

6.4 Derivatives

A *derivative security* (or simply *derivative*) is a security where the value depends explicitly on other variables. In this section, we consider financial instruments whose payoff is explicitly tied to the payoff of other financial securities. This section deals with forward contracts (forwards), futures, and options. Derivatives are traded in a standardized fashion on financial markets (e.g. EUREX, Chicago Board of Trade CBOT) and also in a *over-the-counter* manner, where tailored contracts are sold to investors. The security which determines the value of the derivative is called the *underlying security* (or *underlying*).

The concept of *arbitrage* plays an important role in pricing derivatives. We call the interest rate at which cash can be borrowed or loaned the *risk-free rate of interest*. So, in its simplest form, arbitrage means taking simultaneous positions in different securities guaranteeing a higher profit than the risk-free rate of interest. This would lead to infinite gains by investing infinite borrowed money in the arbitrage opportunity.

6.4.1 Forward Contracts

A forward contract is the simplest example of a derivative. A forward contract is a contract to purchase or sell an asset at a specific price and at a specific time in the future. The *spot price* is the price of the asset for immediate delivery.

The party of the contract who agreed to buy the underlying is said to be *long*. The opposite party who is to sell the underlying takes the *short* position. The price specified in the forward contract is called *delivery price* K. The *forward price* F_t of a forward contract is the delivery price that would apply if the contract would be established at that time. Note that the initial price of a forward contract is zero and therefore no money is transferred at the beginning.

Example 6.2. (Currency Risk)
A Japanese corporation has a contract with a U.S. customer to sell a machine for 1 million dollar in one year. If the exchange rate Yen/Dollar rises in the forthcoming year the Japanese firm will earn less. The Japanese firm can eliminate this risk by entering in forward contract to sell 1 million dollar for fixed amount of Yen in one year.

If we denote the price of the underlying security at maturity by S_T and the forward position by K, the payoff of the long position is

$$S_T - K.$$

This means that the long party of the contract can sell the security at a price of S_T which he bought for K. Of course, the gain of the long position

is the loss of the short position and vice versa. Therefore, the payoff of the short position is

$$K - S_T.$$

Figure 6.4 shows the payoff-diagram of a forward contract. The long position gains on rising exchange rate whereas the short position's earnings rise with declining exchange rate.

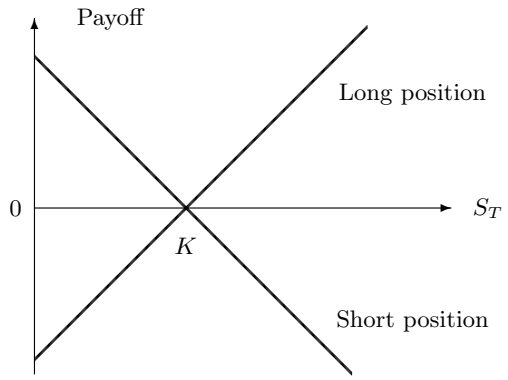

Fig. 6.4. Payoff of a forward contract at maturity where K is the delivery price an S_T the price of the security at maturity

The question which still remains is how to set the delivery price in order to have an initial price of zero. This is the case when there are no arbitrage opportunities.
Delivery price of a forward contract:
Suppose that the current price of an asset is S_0 and assume that the asset can be stored at zero cost. The risk-free interest rate is r. Then, the delivery price in a forward contract with time to delivery T is

$$K = e^{rT} S_0$$

for the continuous case and

$$K = (1 + r)^T S_0$$

for the discrete case.
Proof: The proof is straightforward by showing that there exist arbitrage opportunities otherwise. The proof includes only the continuous-time case. The proof for the discrete-time case in analogous.

Suppose that $K > e^{rT}S_0$. The arbitrage portfolio can now be constructed as follows: Borrow S_0 at rate r and buy the security. At maturity the security can be sold for K and the loan has to be repaid for $e^{rT}S_0$. The guaranteed gain is $K - e^{rT}S_0 > 0$. Therefore, K must not be larger than $e^{rT}S_0$.
Suppose that $K < e^{rT}S_0$. The arbitrage portfolio for this case is the following: Sell the security short and invest the received cash S_0 with the risk-free rate r. At maturity, repurchase the security for K. The gain at maturity is $e^{rT}S_0 - K > 0$. Therefore, K must not be smaller than $e^{rT}S_0$. □

By construction, the value of the contract is zero at the beginning. But as the price of the underlying changes, the value of the forward contract will also change.
Value of a forward contract:
Suppose that the forward price of security S after time t is F_t and that the delivery price of the forward is K. The security can be stored at zero cost. The risk-free interest rate is r. The forward contract has delivery at time T. Then the current value of the forward contract for the long position is

$$f_t = e^{-r(T-t)}(F_t - K)$$

for the continuous-time case and

$$f_t = \frac{1}{(1+r)^{T-t}}(F_t - K)$$

for the discrete-time case.
Obviously the value of the short position is equivalent to the negative value of the long position.
Proof: The proof includes only the continuous-time case, the proof for the discrete-time case in analogous.
Consider a portfolio of a forward with delivery price F_t with short position and one forward with delivery price K with long position at time t. In order to establish such a portfolio, the investor has to lend a cash amount f_t since the forward with delivery price F_t has a value of zero. At time T, the portfolio will be worth $F_t - K$ with certainty and the repayment for the loan will be $f_t e^{r(T-t)}$. There is no arbitrage only if the final value of the loan and the value of the portfolio are identical. □

6.4.2 Futures

Forwards have one major drawback because it is not possible to trade them on an exchange market. In order to make trading possible, the market has to specify certain standards. Therefore, the two parties of a future contract do not need to know each other and the exchange takes care of the risk of default of the counterpart. A future contract has to specify the following:

- The underlying asset, i.e., usually commodities or shares
- The contract size

- Delivery arrangements
- Delivery date
- Quality of delivered goods
- Price quotes
- Daily price movement limits, i.e., prevention of large movements
- Position limits, i.e., maximum of contracts a speculator may hold.

But even by standardizing in the above fashion, there remain two unsolved problems: how to avoid defaults and how to specify the forward price. The setting of the forward price is a very subtle problem since the exchange cannot issue new contracts every time the price of the underlying changes. These two problems can be addressed by margin accounts. Consider the situation where the contract has a forward price of F_0 at the beginning. The next day the forward price is F_1 which is most probably not equal to F_0. The party in the long position then receives $F_1 - F_0$. On the opposite, the short party pays $F_0 - F_1$. In effect, the delivery price in the contract has been changed from F_0 to F_1 and therefore the contract is "renewed" each day. The margin account serves as a guarantee that neither party of the contract defaults on its obligation. On writing the contract, a deposit, known as the initial margin, has to be paid (about 5-10% of the current value of the underlying in the contract). To make sure that the margin account never has a negative balance, a minimum margin, called the maintenance margin, is set to about 75% of the initial margin. If the value of the margin account drops below the maintenance margin, a margin call is issued to the trader and he is requested to top up the margin account to the initial margin.

Since futures are marked to market every day, i.e., the margin account is balanced at the end of every day, the value of the contract is always zero. The only variable price in the contract is the delivery price. As the contract reaches its maturity, the futures price must converge to the spot price (see Figure 6.5). If this is not the case, arbitrage opportunities exist. If the futures

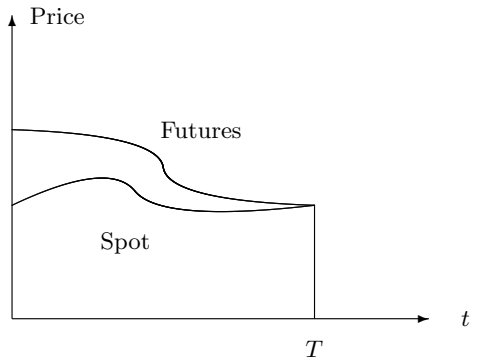

Fig. 6.5. Convergence of futures and spot prices as time approaches maturity

price at maturity would be above the spot price, one could buy the underlying asset at the spot price, take a short position in the future and then settle the contract. Conversely, if the futures price lies below the spot price at maturity one takes a long position in the futures contract and immediately sell the asset in the spot market. Therefore, the futures price has to converge to the spot price.

The question of how to set the futures price is still unanswered. If the risk-free rate is constant and deterministic, the corresponding futures and forward prices are identical. In practice, there are numerous reasons which cause differences between futures and forward prices such as transaction costs, margin accounts, taxes etc. For a more detailed discussion of forwards and futures see [20] and [23].

6.4.3 Options

Option Basics

An option is the right, but not the obligation, to buy (or sell) an asset under specified terms. An option which gives the right to purchase something is called a call option whereas an option which gives the right to sell something is called a put option. An option is a derivative security (asset) whose underlying asset is the asset that can be bought or sold, such as a share. The options on common stocks need specification in order to classify them and price the option. The specification of an option include, firstly, a clear description of what can be bought (for a call) or sold (for a put). Secondly, the exercise price, or strike price (K), must be specified. This is the price at which the asset can be purchased upon exercise of the option. Thirdly, the period of time for which the option is valid must be specified, defined by the expiration date (T). There are two primary conventions regarding acceptable exercise dates before expiration. An American option allows exercise at any time before and including the expiration date. A European option allows exercise only on the expiration date. The terms European and American option classify the different options, but do not imply the location where they are issued. In an option issuing there are two sides involved, the party who grants the option and the party who purchases the option. The party which grants the option is said to write an option. The party which purchases an option faces no risk of loss other than the original purchase premium. However, the party who writes the option, usually a financial intermediary such as a bank or insurance company, may face a large loss, since this party must buy or sell this asset at specified terms if the option is exercised. In the case of an exercised call option, if the writer does not already own the asset, he must purchase it in order to deliver it at the specified strike price, which may be much lower than the current market price.

The Nature of Options

Before we begin to derive the exact pricing formula for options, we start to develop some intuition. Suppose you own a European call option on a

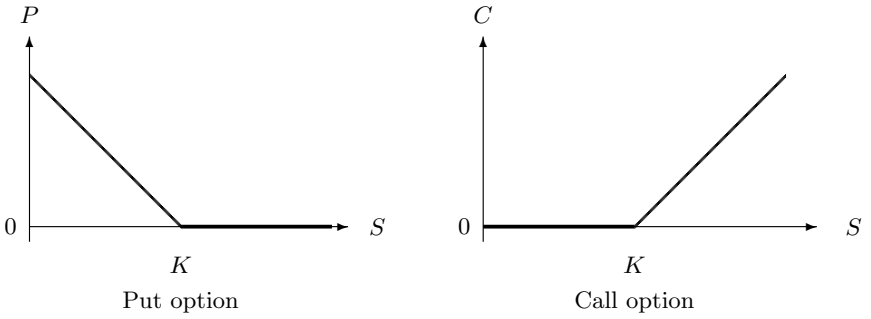

Fig. 6.6. Value of options at expiration date

given stock with strike price K and suppose the the option has reached the expiration time T. What is the value of the option at this time in function of the underlying stock price S? If the stock price S is larger than the strike price K, the value of the call option is $C = S - K$. In the case where $S < K$ you would not exercise the option and thus it has no value, $C = 0$. Therefore, the option payoff at expiration time is

$$C(S,T) = \max(0, S(T) - K).$$

The argument for put options is identical and thus

$$P(S,T) = \max(K - S(T), 0).$$

The value of an option at expiration is graphically depicted in Figure 6.6. In the case where we have not reached the expiration time, the call option value is above its payoff function at expiration time, since the time remaining offers us the possibility that the stock price would further increase and thus the option has a time value. For a call option where $S > K$ we say the option is *in the money*, where $S = K$ we say it is *at the money*, and where $S < K$ we say it is *out of the money*.

In addition to the time value of options, the options has a volatility value. Consider two stocks at the same price and two put option with the same strike price. Stock one is more volatile than stock two. The put option on stock one has a higher value, because the probability that stock one falls below the strike price is higher than for stock two. Since a put option is similar to an

insurance which pays in case of a loss, the insurance premium must go up, if the likelihood of this case gets larger.

Economics of Options

In order to derive pricing concepts for options, we need think about the underlying economics of options and financial markets. Option pricing centers on the concept of arbitrage-free markets. The notion arbitrage is, loosely put, the possibility to earn an abnormal profit with no risks involved. In an arbitrage-free market, one can only obtain a risk-free return at the prevailing risk free interest rate, such as payed by Treasuries bonds. Let us form a portfolio G of a stock at the current price $S(0)$ and a put option P with strike price K. When the market is arbitrage-free, the expected payoff return at expiration r_1 is exactly the risk-free interest rate r, since the combination of a put option and the same underlying stock removes all possibility of a loss. If the payoff is larger than the risk-free rate, we could borrow money (L) at the risk-free rate, and buy the portfolio. We consequently would achieve a profit larger than the risk-free rate at no risk without any capital investments. Because this is such an interesting money machine, we start to buy huge quantities of the portfolio, and this will consequently increase the option price until the payoff at expiration is exactly equal to the risk-free interest rate. A similar argument can be made if the expected payoff is below the risk-free interest rate, because we may short the stock and the option and lend money. The possible scenarios are shown in table below.

Initial Investment	Return	Payoff
$P + S - L = 0$	$r_1 > r$	$E[r_1(P + S) - rL] > 0$
$-P - S + L = 0$	$r_1 < r$	$E[-r_1(P + S) + rL] > 0$
$P + S - L = 0$	$r_1 = r$	$E[r_1(P + S) - rL] = 0$

The same argument can be made to explain the value of options at any other time than expiration. Take a call option some time before expiration which is out of the money. Lets assume the value of the option is zero. We could now obtain infinitely large quantities of this option and face no risk of loss and need no initial capital outlay. The payoff at expiration is either zero or positive and can be larger than the prevailing risk-free interest rate. This would clearly be an arbitrage possibility and thus, the out of the money options must a have a positive time value.

6.4.4 Black-Scholes Formula and PDE

In order to price options correctly, we will derive the famous *Black-Scholes PDE and Formula for Call Options* based on the no arbitrage argument. The derivation of the Black-Scholes Equations can be found in [6], [23], [25], [27], and [14].

Asset Dynamics

The starting point for option pricing is a model of the underlying asset dynamics. For stock option, we need to model the dynamics of stocks.

As outlined before, we use the Geometric Brownian motion model to model stocks for three reasons. First, asset price are always positive, second, stocks behave like bonds with continuous compounding and stochastic returns, and third, stock prices resemble a log-normal distribution. The SDE of the stock price is given by

$$dS(t) = \mu S(t)dt + \sigma S(t)dW \tag{6.2}$$

where μ and σ are constant values. Additionally, we introduce the following more general stock price dynamics

$$dS(t) = \mu(S,t)S(t)dt + \sigma S(t)dW \tag{6.3}$$

where $\mu(S,t)$ is an arbitrary function of S and t, and σ is a constant.

Derivation of the Black-Scholes PDE

The derivation is based on two central arguments. The first argument is that in a perfect market no arbitrage possibilities exist, and the second that the writer of the option should not undertake any risks by writing the option.

Let us call the value of the call option $C(t,S)$, the expiration time T, and the current risk-free interest rate r. The stock price dynamics are assumed to be as in equation (6.2). The call option is a derivative security of the underlying stock S and therefore, it changes its value dC according to Itôs formula

$$dC = \Big(\frac{\partial C}{\partial t} + \frac{\partial C}{\partial S}\mu S + \frac{1}{2}\frac{\partial^2 C}{\partial S^2}\sigma^2 S^2\Big)dt + \frac{\partial C}{\partial S}\sigma S dW \tag{6.4}$$

The writer of the option does not want to undertake any risks by writing the option and therefore forms a portfolio G which is comprised of the amount x of the underlying stock and of the amount y of a bond B, which should match the option value at all times. The portfolio is very often called the replication portfolio. The bond dynamics are

$$dB = rBdt \tag{6.5}$$

The portfolio $G = xS + yB$ should be self-financing, i.e., no money should be needed except for the initial capital outlay. The portfolio dynamics are given by

$$dG = xdS + ydB + dxS + dyB = xdS + ydB \tag{6.6}$$

The term $dxS + dyB$ equals zero, because a change in the amount of stocks and bonds we hold at constant stock and bond prices equals an in- or outflow

of money. Since we require that the portfolio is self-financing, we only allow to change the positions in the stocks and bonds in such way that the change in one asset finance the change in the other asset.

The portfolio dynamics are extended by substituting dS from (6.2) into (6.6). This yields

$$dG = x(\mu S(t)dt + \sigma S(t)dw) + y(rB)dt = (x\mu S + yrB)dt + x\sigma S dW\,. \quad (6.7)$$

Since the portfolio dynamics should match the dynamics of the option, we match the coefficients of dt and dw in (6.7) and (6.4). To do this, we first match the coefficient of dw by setting

$$x = \frac{\partial C}{\partial S}\,. \quad (6.8)$$

Since we want $G = C$ and using $G = xS + yB$ yields

$$G = yB + \frac{\partial C}{\partial S}S = C$$
$$y = \frac{1}{B}\Big(C - \frac{\partial C}{\partial S}S\Big). \quad (6.9)$$

Substituting the (6.8) and (6.9) into (6.7) and matching the coefficient of dt in (6.4) gives

$$\frac{\partial C}{\partial S}\mu S + \frac{1}{B}\Big(C - \frac{\partial C}{\partial S}S\Big)rB = \frac{\partial C}{\partial t} + \frac{\partial C}{\partial S}\mu S + \frac{1}{2}\frac{\partial^2 C}{\partial S^2}\sigma^2 S^2.$$

Finally, we arrive at the celebrated Black-Scholes PDE

$$rC = \frac{\partial C}{\partial t} + \frac{\partial C}{\partial S}rS + \frac{1}{2}\frac{\partial^2 C}{\partial S^2}\sigma^2 S^2. \quad (6.10)$$

The solution of the PDE together with the appropriate boundary conditions gives the pricing of the option. The last step in the argument is to show that (6.10) is arbitrage-free, because before we have only used the argument that the option's writer should not undertake any risks.

Consider the strategy of shorting (writing) one call option and investing the money in the replication portfolio. The initial market price is $C^*(S_0, 0)$ and the theoretical Black-Scholes price is $C(S_0, 0)$. By construction, the replication portfolio always matches the option until the expiration date. If we manage to convince somebody to pay more than the theoretical price, thus $C^*(S_0, 0) > C(S_0, 0)$, we could earn a risk-less profit, because we will invest $C(S_0, 0)$ in the replication portfolio and pocket the initial difference $C^*(S_0, 0) - C(S_0, 0)$ as profit. Since there is no initial capital required and the replication portfolio procedure removes all risks, we would have found an arbitrage possibility. The same argument can be made if the market price is below the theoretical price. Only when $C^*(S_0, 0) = C(S_0, 0)$, any arbitrage possibility is removed. The beauty of the theory is that the Black-Scholes PDE gives a framework to price options and yields a replication procedure that covers all risks for the party that writes the option.

Black-Scholes Formula for European Call Options

Having derived the Black-Scholes PDE, we now present the Black-Scholes formula for European call options.

Before that we take a closer look at the PDE. Let us consider the stock itself. It is (in a trivial way) a derivative of the stock itself, so $C(t,S) = S(t)$ should satisfy the PDE. With this choice we have $\frac{\partial C}{\partial S} = 1$, $\frac{\partial^2 C}{\partial S^2} = 0$, $\frac{\partial C}{\partial t} = 0$ and therefore $rS(t) = rS(t)$. This proves that $C(t,S) = S(t)$ is a possible solution. There are uncountably many more solutions to this PDE.

The Black-Scholes formula for a European call option with strike price K, expiration time T, a non dividend paying stock with current price S, and a constant risk free interest rate r, is given by

$$C(t,S) = S(t)N(d_1) - Ke^{-r(T-t)}N(d_2) \tag{6.11}$$

where $N(d)$ is the cumulative normal probability distribution

$$N(d) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{d} e^{-\frac{y^2}{2}} dy \tag{6.12}$$

and d_1 and d_2 are defined as

$$\begin{aligned} d_1 &= \frac{\ln(\frac{S}{K}) + (r + \frac{\sigma^2}{2})(T-t)}{\sigma\sqrt{T-t}} \\ d_2 &= d_1 - \sigma\sqrt{T-t}\,. \end{aligned}$$

Let us now show that the Black-Scholes formula satisfies the PDE (6.10) and the boundary conditions $C(T,S) = \max(S(T) - K, 0)$. For $t = T$ (expiration time) we get

$$d_1(S,T) = d_2(S,T) = \begin{cases} +\infty & S(T) > K \\ -\infty & S(T) < K \end{cases}.$$

Since $N(\infty) = 1$ and $N(-\infty) = 0$ we obtain

$$C(S,T) = \begin{cases} S - K & S(T) > K \\ 0 & S(T) < K \end{cases}.$$

This proves that (6.11) satisfies the boundary condition. The derivatives of (6.11) given by

$$\begin{aligned} \frac{\partial C}{\partial S} &= N(d_1) \\ \frac{\partial^2 C}{\partial S^2} &= \frac{e^{-\frac{d_1^2}{2}}}{S\sigma\sqrt{2\pi(T-t)}} \\ \frac{\partial C}{\partial t} &= -\frac{e^{-\frac{d_1^2}{2}} S\sigma}{2\sqrt{2\pi(T-t)}} - rKe^{-rT}N(d_2). \end{aligned}$$

Thus,

$$\begin{aligned}\frac{\partial C}{\partial t}+\frac{\partial C}{\partial S}rS+\frac{1}{2}\frac{\partial^2 C}{\partial S^2}\sigma^2S^2 &= -\frac{e^{-\frac{d_1^2}{2}}S\sigma}{2\sqrt{2\pi(T-t)}}-rKe^{-rT}N(d_2)\\ &\quad +N(d_1)rS+\frac{e^{-\frac{d_1^2}{2}}}{2S\sigma\sqrt{2\pi(T-t)}}\sigma^2S^2\\ &= r\Big(S(t)N(d_1)-Ke^{-r(T-t)}N(d_2)\Big)\\ &= rC(t,S).\end{aligned}$$

This shows that (6.11) satisfies the Black-Scholes PDE. The Black-Scholes PDE and the call option formula is valid not only when we model the stock price dynamics with a geometric Brownian motion (6.2), but also with the more general dynamics shown in (6.3). Since the drift term does not enter the PDE, it does not affect the pricing of options. The replication procedure remains the same as well. For a detailed discussion of this property, the reader may refer to [9].

6.4.5 Black-Scholes Formula for European Put Options

The formula for European put options can be easily derived from the formula for European call options. To derive the formula, we need to use a simple theoretical relationship between the prices of the corresponding puts and calls. If you buy a call and sell a put with the same strike price, the portfolio behaves almost like the stock itself. The payoff at expiration time is $\max(S(T)-K,0)-\max(K-S(T),0)=S(T)-K$. The difference to the stock price is the strike price K. By lending $e^{-r(T-t)}K$, we obtain a payoff of K at expiration time and the portfolio of the put and the call and the credit resembles exactly the stock. Thus the call-put-parity is $C-P+e^{-r(T-t)}K=S$.

We now use the so-called put-call-parity and the Black-Scholes call option formula to derive the European put option formula.

$$\begin{aligned}C-P+e^{-r(T-t)}K &= S\\ S(t)N(d_1)-Ke^{-r(T-t)}N(d_2)-P(t,S)+e^{-r(T-t)}K &= S(t)\end{aligned}$$

$$P(t,S)=S(t)(N(d_1)-1)-Ke^{-r(T-t)}(1-N(d_2)).$$

Since $(N(x)-1)=N(-x)$ we can state the formula for European put options

$$P(t,S)=Ke^{-r(T-t)}N(-d_2)-S(t)N(-d_1).$$

American Options

In the option contracts we have discussed so far, the holder may exercise the option at a certain time specified in the contract. Often however, the holder is

given the opportunity to exercise early which means that the option may be exercised not only on a given date but also any at time before. Such contracts are known as American options. A mixture between American and European options are options which can be exercised at a fixed number of dates before the expiration date. Those options (you might guess it) are called "Bermuda" options.

American options or any other "Exotic" options, where the payoff depends not only of the value of the underlying security at expiration date but on the path of the asset, are a lot more difficult to evaluate. The valuation of American options usually are based on a worst case principle, where it is assumed that the option will be exercised at a time at which this is most profitable to the holder, which represents a worst case for the writer. We will now use the example of an American put option to illustrate the pricing of American options. For American options the Black-Scholes PDE is replaced by an inequality

$$\frac{\partial P}{\partial t} + \frac{\partial P}{\partial S} rS + \frac{1}{2}\frac{\partial^2 P}{\partial S^2}\sigma^2 S^2 - rP \leq 0 \tag{6.13}$$

in which equality holds if the option is not exercised, i.e., if its value exceeds the revenue of exercising. For the put option, this is expressed by the inequality

$$P(t, S) > \max(K - S(t), 0).$$

For the PDE (6.13) the condition on maturity (expiration) holds:

$$P(T, S) = \max(K - S(T), 0).$$

In addition we impose the boundary conditions on the PDE. The first boundary condition states that the put has no value for arbitrarily large stock prices

$$\lim_{S \to \infty} P(T, S) = 0.$$

and the second boubdary condition states that the value of the put option for $S(t) = 0$ is equal to the discounted strike price

$$P(T, 0) = e^{-r(T-t)} K\,.$$

The Black-Scholes PDE (6.13) together with the boundary conditions, the terminal value, and the early exercise inequality yield no analytical solution. The PDE needs to be solved numerically, where one of the popular techniques is the transformation into a linear complimentary system. The solution procedure is given in detail in ([33]).

6.4.6 General Option Pricing

An similar PDE for pricing can be derived for any single factor stock price process. If we assume that the underlying asset follows an SDE as given by

$$S(t) = \mu(t, S)dt + \sigma(t, S)dW$$

the PDE for pricing any possible option $C(t, S)$ is given by

$$rC = \frac{\partial C}{\partial t} + \frac{\partial C}{\partial S} rS + \frac{1}{2} \frac{\partial^2 C}{\partial S^2} \sigma(t, S)^2 \,. \tag{6.14}$$

Again it is noteworthy, that the drift does not enter the PDE and that the diffusion term changes the PDE. The derivation is analogous to the derivation of the Black-Scholes PDE. This equation (6.14) applies only when the underlying stock price is modelled by a scalar SDE. In the case that a system of SDE models the underlying stock, a multi dimensional form of (6.14) can be derived. The reader may refer to [25] or [14].

References

1. N. U. Ahmed. *Linear and Nonlinear Filtering for Scientists and Engineers*. World Scientific, Singapore, 1998.
2. B. D. O. Anderson and J. B. Moore. *Optimal Filtering*. Prentice-Hall, Englewood Cliffs, NJ, 1979.
3. L. Arnold. *Stochastische Differentialgleichungen: Theorie und Anwendung*. Oldenbourg, München, 1973.
4. M. Athans and H. P. Geering. Necessary and sufficient conditions for non-scalar-valued functions to attain extrema. *IEEE Transactions on Automatic Control*, 18:132–139, 1973.
5. N. H. Bingham and R. Kiesel. *Risk Neutral Valuation*. Springer, London, 1998.
6. T. Björk. *Arbitrage Theory in Continuous Time*. Oxford University Press, Oxford, 1998.
7. L. Breiman. *Probability*. SIAM, Philadelphia, PA, reprinted edition, 1992.
8. A. E. Bryson and Y.-C. Ho. *Applied Optimal Control: Optimization, Estimation, and Control*. Hemisphere, Washington, DC, 1975.
9. J. Y. Campbell, A. W. Lo, and A. C. MacKinlay. *The Econometrics of Financial Markets*. Princeton University Press, Princeton, NJ, second edition, 1997.
10. J. Y. Campbell and L. M. Viceira. *Strategic Asset Allocation: Portfolio Choice for Long-Term Investors*. Oxford University Press, Oxford, 2002.
11. S. Cyganowski, P. Kloeden, and J. Ombach. *From Elementary Probability to Stochastic Differential Equations with MAPLE*. Springer, Berlin, 2002.
12. J. E. Dennis and R. B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. SIAM, Philadelphia, PA, 1996.
13. J. L. Doob. *Stochastic Processes*. Wiley, New York, NY, 1953.
14. D. Duffie. *Dynamic Asset Pricing Theory*. Princeton University Press, Princeton, NJ, second edition, 1996.
15. W. H. Fleming and R. W. Rishel. *Deterministic and Stochastic Optimal Control*. Springer, New York, NY, 1975.
16. H. P. Geering. *Regelungstechnik*. Springer, Berlin, sixth edition, 2003.
17. H. P. Geering. *Optimal Control*. Springer, Berlin, 2007.
18. J. D. Hamilton. *Time Series Analysis*. Princeton University Press, Princeton, NJ, 1994.
19. A. C. Harvey. *Zeitreihenmodelle*. Oldenbourg, München, 1995.

20. J. C. Hull. *Options, Futures, and Other Derivatives*. Pearson Prentice Hall, Uppersaddle River, NJ, sixth edition, 1999.
21. A. H. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, New York, NY, 1970.
22. P. Kellerhals. *Financial Pricing Models in Continuous Time and Kalman Filtering*. Springer, Berlin, 2001.
23. D. G. Luenberger. *Investment Science*. Oxford University Press, Oxford, 1998.
24. A. G. Malliaris and W. A. Brock. *Stochastic Methods in Economics and Finance*. Elsevier Science, Amsterdam, 1982.
25. R. C. Merton. *Continuous-Time Finance*. Blackwell, Malden, MA, second edition, 1992.
26. T. Mikosch. *Elementary Stochastic Calculus with Finance in View*. World Scientific, Singapore, 1998.
27. S. N. Neftci. *An Introduction to the Mathematics of Financial Derivatives*. Academic Press, San Diego, CA, second edition, 2000.
28. B. Øksendal. *Stochastic Differential Equations*. Springer, Berlin, fifth edition, 1998.
29. H. H. Panjer, editor. *Financial Economics*. The Actuarial Foundation, Schaumburg, IL, 1998.
30. W. Paul and J. Baschnagel. *Stochastic Processes: From Physics to Finance*. Springer, Berlin, 2000.
31. F. K. Reilly and K. C. Brown. *Investment Analysis and Portfolio Management*. Thomson South-Western, Mason, OH, eigth edition, 2006.
32. H. Schlitt. *Systemtheorie für regellose Vorgänge*. Springer, Berlin, 1960.
33. H. Schumacher, B. Hanzon, and M. Vellekoop. *Finance for Control Engineers: Tutorial Workshop*. European Control Conference, Porto, 2001.
34. W. F. Sharpe, G. J. Alexander, and J. V. Bailey. *Investments*. Prentice Hall, Uppersaddle River, NJ, sixth edition, 1999.
35. J.-J. E. Slotine and W. Li. *Applied Nonlinear Control*. Prentice Hall, Englewood Cliffs, NJ, 1991.
36. D. Williams. *Probability with Martingales*. Cambridge University Press, Cambridge, 1991.
37. J. Yong and X. Y. Zhou. *Stochastic Controls: Hamiltonian Systems and HJB Equations*. Springer, New York, NY, 1999.