Stochastic and Partial Differential Equations with Adapted Numerics [1]

Jonathan Goodman Kyoung-Sook Moon
Anders Szepessy Raúl Tempone Georgios Zouraris

January 20, 2004

[1]This is a draft. Comments and improvements are welcome.

Contents

Chapter 1

Introduction and Motivating Examples

The goal of this course is to give useful understanding to solve problems formulated by stochastic or partial differential equations models in science, engineering and mathematical finance. Typically, these problems require numerical methods to obtain a solution and therefore the course focuses on basic understanding of stochastic and partial differential equations to construct reliable and efficient computational methods.

In particular, we will study the amount of computational work for alternative numerical methods to solve a problem with a given accuracy. The optimal method clearly minimizes the work for given accuracy. Therefore it is valuable to know something about accuracy and work for different numerical methods, which lead us to error estimates and convergence results.

1.1 Noisy Evolution of Stock Values

Let us consider a stock value denoted by the time dependent function $S(t)$. To begin our discussion, assume that $S(t)$ satisfies the differential equation

$$\frac{dS}{dt} = a(t)S(t),$$

which has the solution

$$S(t) = e^{\int_0^t a(u)du} S(0).$$

Our aim is to introduce some kind of noise in the above simple model of the form $a(t) = r(t) + "noise"$, taking into account that we do not know precisely how the evolution will be. An example of a "noisy" model we shall consider is the stochastic differential equation

$$dS(t) = r(t)S(t)dt + \sigma S(t)dW(t), \tag{1.1}$$

where $dW(t)$ will introduce noise in the evolution. To seek a solution for the above, the starting point will be the discretization

$$S_{n+1} - S_n = r_n S_n \Delta t_n + \sigma_n S_n \Delta W_n, \tag{1.2}$$

where ΔW_n are independent normally distributed random variables with zero mean and variance Δt_n, i.e. $E[\Delta W_n] = 0$ and $Var[\Delta W_n] = \Delta t_n = t_{n+1} - t_n$. As will be seen later on, (1.1) may have more than one possible interpretation, and the characterization of a solution will be intrinsically associated with the numerical discretization used to solve it.

We shall consider, among others, applications to option pricing problems. An European call option is a contract which gives the right, but not the obligation, to buy a stock for a fixed price K at a fixed future time T. The celebrated Black-Scholes model for the value $f : (0, T) \times (0, \infty) \to \mathbb{R}$ of an option is the partial differential equation

$$\begin{aligned} \partial_t f + rs\partial_s f + \frac{\sigma^2 s^2}{2}\partial_s^2 f &= rf, \ 0 < t < T, \\ f(s, T) &= \max(s - K, 0), \end{aligned} \tag{1.3}$$

where the constants r and σ denote the riskless interest rate and the volatility respectively. If the underlying stock value S is modeled by the stochastic differential equation (1.1) satisfying $S(t) = s$, the Feynmann-Kač formula gives the alternative probability representation of the option price

$$f(s, t) = E[e^{-r(T-t)} \max(S(T) - K, 0))|S(t) = s], \tag{1.4}$$

which connects the solution of a partial differential equation with the expected value of the solution of a stochastic differential equation. Although explicit exact solutions can be found in particular cases, our emphasis will be on general problems and numerical solutions. Those can arise from discretization of (1.3), by finite difference or finite elements methods, or from Monte Carlo methods based on statistical sampling of (1.4), with a discretization (1.2). Finite difference and finite element methods lead to a discrete system of equations substituting derivatives for difference quotients, (e.g.)

$f_t \approx \frac{f(t_{n+1})-f(t_n)}{\Delta t}$; the Monte Carlo method discretizes a probability space, substituting expected values by averages of finite samples, e.g. $\{S(T,\omega_j)\}_{j=1}^M$ and $f(s,t) \approx \sum_{j=1}^M \frac{e^{-r(T-t)}\max(S(T,\omega_j)-K,0)}{M}$. Which method is best? The solution depends on the problem to solve and we will carefully study qualitative properties of the numerical methods to understand the answer.

1.2 Porous Media Flow

An other motivation for stochastic differential equations is provided by *porous media flow*. In this case the uncertainty comes from the media where the flow takes place. The governing equations are the continuity equation of an incompressible flow

$$div(V) = 0, \tag{1.5}$$

and Darcy's law

$$V = -K\nabla P, \tag{1.6}$$

where V represents the flow velocity and P is the pressure field. The function K, the so called conductivity of the material, is the source of randomness, since in practical cases, it is not precisely known. We would like to study the concentration C of an inert pollutant carried by the flow V, satisfying the convection equation

$$\partial_t C + V \cdot \nabla C = 0.$$

The variation of K is, via Darcy's laws (1.6), important to determine properties of the concentration C. One way to determine the flow velocity is to solve the pressure equation

$$div(K\nabla P) = 0, \tag{1.7}$$

in a domain with given values of the pressure on the boundary of this domain. Assume that the flow is two dimensional with $V = (1, \hat{V})$, where $\hat{V}(x)$ is stochastic with mean zero, i.e. $E[\hat{V}] = 0$. Thus,

$$\partial_t C + \partial_x C + \hat{V}\partial_y C = 0.$$

Let us define $\bar{C}$ as the solution of $\partial_t \bar{C} + \partial_x \bar{C} = 0$. We wonder if $\bar{C}$ is the expected value of C, i.e. is $\bar{C} \stackrel{?}{=} E[C]$? The answer is in general no. The difference comes from the expected value

$$E[\hat{V}\partial_y C] \neq E[\hat{V}]E[\partial_y C] = 0.$$

The desired averaged quantity $\tilde{C} = E[C]$ is an example of turbulent diffusion and in the simple case $\hat{V}(x)dx = dW(x)$ (cf. (1.1)) it will satisfy a convection diffusion equation of the form

$$\partial_t \tilde{C} + \partial_x \tilde{C} = \frac{1}{2}\partial_{yy}\tilde{C},$$

which is related to the Feynman-Kač formula (1.4). We will develop efficient numerical methods for more general stochastic velocities.

1.3 Optimal Control of Investments

Suppose that we invest in a risky asset, whose value $S(t)$ evolves according to the stochastic differential equation $dS(t) = \mu S(t)dt + \sigma S(t)dW(t)$, and in a riskless asset $Q(t)$ that evolves with $dQ(t) = rQ(t)dt$, $r < \mu$. Our total wealth is then $X(t) = Q(t) + S(t)$ and the goal is to determine an optimal instantaneous policy of investment in order to maximize the expected value of our wealth at a given final time T. Let $\alpha(t)$ be defined by $\alpha(t)X(t) = S(t)$, so that $(1 - \alpha(t))X(t) = Q(t)$ with $\alpha \in [0, 1]$. Then our optimal control problem can be stated as

$$\max_{\alpha} E[g(X(T))|X(t) = x] \equiv u(t, x),$$

where g is a given function. How can we determine α? The solution of this problem can be obtained by means of a Hamilton Jacobi equation, which is in general a nonlinear partial differential equation of the form

$$u_t + H(u, u_x, u_{xx}) = 0.$$

Part of our work is to study the theory of Hamilton Jacobi equations and numerical methods for control problems to determine the Hamiltonian H and the control α.

Chapter 2

Stochastic Integrals

This chapter introduces stochastic integrals, which will be the basis for stochastic differential equations in the next chapter. Here we construct approximations of stochastic integrals and prove an error estimate. The error estimate is then used to establish existence and uniqueness of stochastic integrals, which has the interesting ingredient of intrinsic dependence on the numerical approximation due to infinite variation. Let us first recall the basic definitions of probability we will use.

2.1 Probability Background

A probability space is a triple $(\Omega, \mathcal{F}, P)$, where Ω is the set of outcomes, $\mathcal{F}$ is the set of events and $P : \mathcal{F} \to [0, 1]$ is a function that assigns probabilities to events satisfying the following definitions.

Definition 2.1 If Ω is a given non empty set, then a *σ-algebra* $\mathcal{F}$ on Ω is a collection $\mathcal{F}$ of subsets of Ω that satisfy:

(1) $\Omega \in \mathcal{F}$;

(2) $F \in \mathcal{F} \Rightarrow F^c \in \mathcal{F}$, where $F^c = \Omega - F$ is the complement set of F in Ω; and

(3) $F_1, F_2, \ldots \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{+\infty} F_i \in \mathcal{F}$.

Definition 2.2 A probability measure on $(\Omega, \mathcal{F})$ is a set function $P : \mathcal{F} \to [0, 1]$ such that:

(1) $P(\emptyset) = 0,\ P(\Omega) = 1$; and

(2) If $A_1, A_2, \ldots \in \mathcal{F}$ are mutually disjoint sets then

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

Definition 2.3 A random variable X, in the probability space $(\Omega, \mathcal{F}, P)$, is a function $X : \Omega \to \mathbb{R}^d$ such that the inverse image $X^{-1}(A) \equiv \{\omega \in \Omega : X(\omega) \in A\} \in \mathcal{F}$, for all open subsets A of $\mathbb{R}^d$.

Definition 2.4 [Independence of stochastic variables] Two sets $A, B \in \mathcal{F}$ are said to be independent if

$$P(A \cap B) = P(A)P(B).$$

Two independent random variables X, Y in $\mathbb{R}^d$ are independent if

$$X^{-1}(A) \text{ and } Y^{-1}(B) \text{ are independent for all open sets } A, B \subseteq \mathbb{R}^d.$$

Definition 2.5 A stochastic process $X : [0, T] \times \Omega \to \mathbb{R}^d$ in the probability space $(\Omega, \mathcal{F}, P)$ is a function such that $X(t, \cdot)$ is a random variable in $(\Omega, \mathcal{F}, P)$ for all $t \in (0, T)$. We will often write $X(t) \equiv X(t, \cdot)$.

The t variable will usually be associated with the notion of time.

Definition 2.6 Let $X : \Omega \to \mathbb{R}$ be a random variable and suppose that the density function

$$p'(x) = \frac{P(X \in dx)}{dx}$$

is integrable. The expected value of X is then defined by the integral

$$E[X] = \int_{-\infty}^{\infty} x p'(x) dx, \tag{2.1}$$

which also can be written

$$E[X] = \int_{-\infty}^{\infty} x dp(x). \tag{2.2}$$

The last integral makes sense also in general when the density function is a measure, e.g. by successive approximation with random variables possessing integrable densities. A point mass, i.e. a Dirac delta measure, is an example of a measure.

Exercise 2.7 Show that if X, Y are independent random variables then

$$E[XY] = E[X]E[Y].$$

2.2 Brownian Motion

As a first example of a stochastic process, let us introduce

Definition 2.8 [The Wiener process] The one-dimensional *Wiener process* $W : [0,\infty) \times \Omega \to \mathbb{R}$, also known as the Brownian motion, has the following properties:

(1) with probability 1, the mapping $t \mapsto W(t)$ is continuous and $W(0) = 0$;

(2) if $0 = t_0 < t_1 < \ldots < t_N = T$, then the increments

$$W(t_N) - W(t_{N-1}), \ldots, W(t_1) - W(t_0)$$

are *independent*; and

(3) for all $t > s$ the increment $W(t) - W(s)$ has the *normal* distribution, with $E[W(t) - W(s)] = 0$ and $E[(W(t) - W(s))^2] = t - s$, i.e.

$$P(W(t) - W(s) \in \Gamma) = \int_\Gamma \frac{e^{\frac{-y^2}{2(t-s)}}}{\sqrt{2\pi(t-s)}} dy, \quad \Gamma \subset \mathbb{R}.$$

Does there exists a Wiener process and how to construct W if it does? In computations we will only need to determine W at finitely many time steps $\{t_n : n = 0, \ldots, N\}$ of the form $0 = t_0 < t_1 < \ldots < t_N = T$. The definition then shows how to generate $W(t_n)$ by a sum of independent normal distributed random variables, see Example 2.18 for computational methods to generate independent normal distributed random variables. These independent increments will be used with the notation $\Delta W_n = W(t_{n+1}) - W(t_n)$. Observe, by Properties 1 and 3, that for fixed time t the Brownian motion $W(t)$ is itself a normal distributed random variable. To generate W for all $t \in \mathbb{R}$ is computationally infeasible, since it seems to require infinite computational work. Example 2.18 shows the existence of W by proving uniform convergence of successive continuous piecewise linear approximations. The approximations are based on an expansion in the orthogonal $L^2(0,T)$ Haar-wavelet basis, which will be further studied in Section 9.2 to develop fast computational methods for the porous media problem of Section 1.2.

2.3 Approximation and Definition of Stochastic Integrals

Remark 2.9 [Questions on the definition of a stochastic integral] Let us consider the problem of finding a reasonable definition for the stochastic integral $\int_0^T W(t)dW(t)$, where $W(t)$ is the Wiener process. As a first step, let us discretize the integral by means of the *forward Euler* discretization

$$\sum_{n=0}^{N-1} W(t_n) \underbrace{(W(t_{n+1}) - W(t_n))}_{=\Delta W_n}.$$

Taking expected values we obtain by Property 2 of Definition 2.8

$$E[\sum_{n=0}^{N-1} W(t_n)\Delta W_n] = \sum_{n=0}^{N-1} E[W(t_n)\Delta W_n] = \sum_{n=0}^{N-1} E[W(t_n)] \underbrace{E[\Delta W_n]}_{=0} = 0.$$

Now let us use instead the *backward Euler* discretization

$$\sum_{n=0}^{N-1} W(t_{n+1})\Delta W_n.$$

Taking expected values yields a different result:

$$\sum_{n=0}^{N-1} E[W(t_{n+1})\Delta W_n] = \sum_{n=0}^{N-1} E[W(t_n)\Delta W_n] + E[(\Delta W_n)^2] = \sum_{n=0}^{N-1} \Delta t = T \neq 0.$$

Moreover, if we use the *trapezoidal* method the result is

$$\begin{aligned} \sum_{n=0}^{N-1} E\left[\frac{W(t_{n+1}) + W(t_n)}{2}\Delta W_n\right] &= \sum_{n=0}^{N-1} E[W(t_n)\Delta W_n] + E[(\Delta W_n)^2/2] \\ &= \sum_{n=0}^{N-1} \frac{\Delta t}{2} = T/2 \neq 0. \end{aligned}$$

□

Remark 2.9 shows that we need more information to define the stochastic integral $\int_0^t W(s)dW(s)$ than to define a deterministic integral. We must decide if the solution we seek is the limit of the forward Euler method. In fact, limits of the forward Euler define the so called *Itô integral*, while the trapezoidal method yields the so called *Stratonovich integral*. It is useful to define the class of stochastic processes which can be Itô integrated. We shall restrict us to a class that allows computable quantities and gives convergence rates of numerical approximations. For simplicity, we begin with Lipschitz continuous functions in $\mathbb{R}$ which satisfy (2.3) below. The next theorem shows that once the discretization method is fixed to be the forward Euler method, the discretizations converge in L^2. Therefore the limit of forward Euler discretizations is well defined, i.e. the limit does not depend on the sequence of time partitions, and consequently the limit can be used to define the Itô integral.

Theorem 2.10 *Suppose there exist a positive constant C such that $f : [0,T] \times \mathbb{R} \to \mathbb{R}$ satisfies*

$$|f(t+\Delta t, W+\Delta W) - f(t,W)| \leq C(\Delta t + |\Delta W|). \tag{2.3}$$

Consider two different partitions of the time interval $[0,T]$

$$\begin{aligned} &\{\bar{t}_n\}_{n=0}^{\bar{N}}, \ \bar{t}_0 = 0, \bar{t}_{\bar{N}} = T, \\ &\left\{\bar{\bar{t}}_m\right\}_{m=0}^{\bar{\bar{N}}} \ \bar{\bar{t}}_0 = 0, \bar{\bar{t}}_{\bar{\bar{N}}} = T, \end{aligned}$$

with the corresponding forward Euler approximations

$$\bar{I} = \sum_{n=0}^{\bar{N}-1} f(\bar{t}_n, W(\bar{t}_n))(W(\bar{t}_{n+1}) - W(\bar{t}_n)), \tag{2.4}$$

$$\bar{\bar{I}} = \sum_{m=0}^{\bar{\bar{N}}-1} f(\bar{\bar{t}}_m, W(\bar{\bar{t}}_m))(W(\bar{\bar{t}}_{n+1}) - W(\bar{\bar{t}}_n)). \tag{2.5}$$

Let the maximum time step Δt_{max} be

$$\Delta t_{max} = \max\left[\max_{0\leq n\leq \bar{N}-1} \bar{t}_{n+1} - \bar{t}_n, \max_{0\leq m\leq \bar{\bar{N}}-1} \bar{\bar{t}}_{m+1} - \bar{\bar{t}}_m\right].$$

Then

$$E[(\bar{I} - \bar{\bar{I}})^2] \leq \mathcal{O}(\Delta t_{max}). \tag{2.6}$$

Proof. It is useful to introduce the finer grid made of the union of the nodes on the two grids

$$\{t_k\} \equiv \{\bar{t}_n\} \cup \{\bar{\bar{t}}_m\}.$$

Then in that grid we can write

$$\bar{I} - \bar{\bar{I}} = \sum_k \Delta f_k \Delta W_k,$$

where $\Delta f_k = f(\bar{t}_n, W(\bar{t}_n)) - f(\bar{\bar{t}}_m, W(\bar{\bar{t}}_m))$, $\Delta W_k = W(t_{k+1}) - W(t_k)$ and the indices m, n satisfy $t_k \in [\bar{\bar{t}}_m, \bar{\bar{t}}_{m+1})$ and $t_k \in [\bar{t}_n, \bar{t}_{n+1})$, as depicted in Figure 2.3.

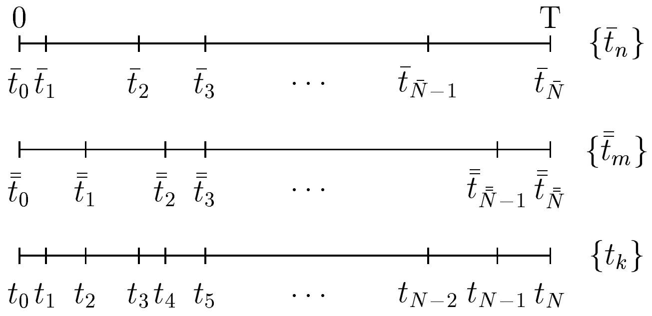


Figure 2.1: Mesh points used in the proof.

Therefore,

$$\begin{aligned} E[(\bar{I} - \bar{\bar{I}})^2] &= E[\sum_{k,l} \Delta f_k \Delta f_l \Delta W_l \Delta W_k] \\ &= 2 \sum_{k>l} \underbrace{E[\Delta f_k \Delta f_l \Delta W_l \Delta W_k]}_{=E[\Delta f_k \Delta f_l \Delta W_l] E[\Delta W_k]=0} + \sum_k E[(\Delta f_k)^2 (\Delta W_k)^2] \\ &= \sum_k E[(\Delta f_k)^2] E[(\Delta W_k)^2] = \sum_k E[(\Delta f_k)^2] \Delta t_k. \end{aligned} \tag{2.7}$$

Taking squares in (2.3) we arrive at $|\Delta f_k|^2 \leq 2C^2((\Delta' t_k)^2 + (\Delta' W_k)^2)$ where $\Delta' t_k = \bar{t}_n - \bar{\bar{t}}_m \leq \Delta t_{max}$ and $\Delta' W_k = W(\bar{t}_n) - W(\bar{\bar{t}}_m)$, using also the standard inequality $(a+b)^2 \le 2(a^2+b^2)$. Substituting this in (2.7) proves the theorem

$$\begin{aligned} E[(\bar I - \bar{\bar I})^2] &\le \sum_k 2C^2 \left((\Delta' t_k)^2 + \underbrace{E[(\Delta' W_k)^2]}_{=\Delta' t_k} \right) \Delta t_k \\ &\le 2C^2\, T(\Delta t_{max}^2 + \Delta t_{max}). \end{aligned} \tag{2.8}$$

□

Thus, the sequence of approximations $I_{\Delta t}$ is a Cauchy sequence in the Hilbert space of random variables generated by the norm $\|I_{\Delta t}\|_{L^2} \equiv \sqrt{E[I_{\Delta t}^2]}$ and the scalar product $(X,Y) \equiv E[XY]$. The limit I of this Cauchy sequence defines the Itô integral

$$\sum_i f_i \Delta W_i \overset{L^2}{\to} I \equiv \int_0^T f(s, W(s)) dW(s).$$

Remark 2.11 [Accuracy of strong convergence] If $f(t, W(t)) = \bar f(t)$ is independent of $W(t)$ we have first order convergence $\sqrt{E[(\bar I - \bar{\bar I})^2]} = \mathcal{O}(\Delta t_{max})$, whereas if $f(t, W(t))$ depends on $W(t)$ we only obtain one half order convergence $\sqrt{E[(\bar I - \bar{\bar I})^2]} = \mathcal{O}(\sqrt{\Delta t_{max}})$. The constant C in (2.3) and (2.9) measures the computational work to approximate the integral with the Euler method: an approximation error $\epsilon \le 2C\sqrt{T\Delta t}$, with uniform step size, requires by (2.8) the computational work corresponding to $N = T/\Delta t \ge 4T^2C^2/\epsilon^2$ steps.

Exercise 2.12 Use the forward Euler discretization to show that

$$\int_0^T s\; dW(s) = TW(T) - \int_0^T W(s) ds$$

Definition 2.13 A process $f : [0,T] \times \Omega \to \mathbb{R}$ is *adapted* if $f(t,\cdot)$ only depends on events which are generated by $W(s)$, $s \le t$.

Remark 2.14 [Extension to adapted Itô integration] Itô integrals can be extended to adapted processes. Assume $f : [0,T] \times \Omega \to \mathbb{R}$ is adapted and that there is a constant C such that

$$\sqrt{E[|f(t+\Delta t, \omega) - f(t,\omega)|^2]} \le C\sqrt{\Delta t}. \tag{2.9}$$

Then the proof of Theorem 2.10 shows that (2.4-2.6) still hold.

Theorem 2.15 (Basic properties of Itô integrals)
Suppose that $f, g : [0,T]\times\Omega \to \mathbb{R}$ are Itô integrable, e.g. adapted and satifying (2.9), and that c_1, c_2 are constants in $\mathbb{R}$. Then:

(1) $\int_0^T (c_1 f(s,\cdot)+c_2 g(s,\cdot))dW(s) = c_1\int_0^T f(s,\cdot)dW(s)+c_2\int_0^T g(s,\cdot)dW(s)$.

(2) $E\left[\int_0^T f(s,\cdot)dW(s)\right] = 0$.

(3)

$$E\left[(\int_0^T f(s,\cdot)dW(s))(\int_0^T g(s,\cdot)dW(s))\right] = \int_0^T E\left[f(s,\cdot)g(s,\cdot)\right]ds.$$

Proof. To verify Property 2, show first that for an Euler discretization

$$E[\sum_{n=0}^{N-1} f(t_n,\cdot)\Delta W_n] = \sum_{n=0}^{N-1} E[f(t_n,\cdot)]E[\Delta W_n] = 0.$$

Then the convergence result (2.6) implies that

$$\begin{aligned}
|E[\int_0^T f(t,\cdot)dW(t)]| &= |E[\int_0^T f(t,\cdot)dW(t) - \sum_{n=0}^{N-1} f(t_n,\cdot)\Delta W_n] + E[\sum_{n=0}^{N-1} f(t_n,\cdot)\Delta W_n]| \\
&\le \sqrt{E[\left(\int_0^T f(t,\cdot)dW(t) - \sum_{n=0}^{N-1} f(t_n,\cdot)\Delta W_n\right)^2]} \to 0.
\end{aligned}$$

Property 1 and 3 can be verified analogously. □

Exercise 2.16 Use the forward Euler discretization to show that

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

(2) Property 1 and 3 in Theorem 2.15 hold.

Exercise 2.17 Consider the Ornstein-Uhlenbeck process defined by

$$X(t) = X_\infty + e^{-at}(X(0) - X_\infty) + b\int_0^t e^{-a(t-s)}dW(s), \qquad (2.10)$$

where X_∞, a and b are given real numbers. Use the properties of the Itô integral to compute $E[X(t)]$, $Var[X(t)]$, $\lim_{t\to\infty} E[X(t)]$ and $\lim_{t\to\infty} Var[X(t)]$. Can you give an intuitive interpretation of the result?

Example 2.18 [Existence of a Wiener process] To construct a Wiener process on the time interval $[0,T]$, define the Haar-functions H_i by $H_0(t) \equiv 1$ and for $2^n \le i < 2^{n+1}$ and $n = 0,1,2\ldots$, by

$$H_i(t) = \begin{cases} T^{-1/2}2^{n/2} & \text{if } (i-2^n)2^{-n} \le t/T < (i+0.5-2^n)2^{-n}, \\ -T^{-1/2}2^{n/2} & \text{if } (i+0.5-2^n)2^{-n} \le t/T < (i+1-2^n)2^{-n}, \\ 0 & \text{otherwise.} \end{cases} \tag{2.11}$$

Then $\{H_i\}$ is an orthonormal basis of $L^2(0,T)$, (why?). Define the continuous piecewise linear function $W^{(m)} : [0,T] \to \mathbb{R}$ by

$$W^{(m)}(t) = \sum_{i=1}^{m} \xi_i S_i(t), \tag{2.12}$$

where $\xi_i,\ i = 1,\ldots,m$ are independent random variables with the normal distribution $N(0,1)$ and

$$S_i(t) = \int_0^t H_i(s)ds = \int_0^T 1_{(0,t)}(s)H_i(s)ds,$$

$$1_{(0,t)}(s) = \begin{cases} 1 & \text{if } s \in (0,t), \\ 0 & \text{otherwise.} \end{cases}$$

The functions S_i are small "hat"-functions with a maximum value $T^{-1/2}2^{-(n+2)/2}$ and zero outside an interval of length $T2^{-n}$. Let us postpone the proof that $W^{(m)}$ converge uniformly and first assume this. Then the limit $W(t) = \sum_{i=1}^{\infty} \xi_i S_i(t)$ is continuous. To verify that the limit W is a Wiener process, we first observe that $W(t)$ is a sum of normal distributed variables so that $W(t)$ is also normal distributed. It remains to verify that the increments ΔW_n and ΔW_m are independent, for $n \ne m$, and $E[(\Delta W_n)^2] = \Delta t_n$. Parseval's equality shows the independence and the correct variance

$$\begin{aligned}
&E[\Delta W_n \Delta W_m] \\
&\quad= E[\sum_{i,j} \xi_i\xi_j (S_i(t_{n+1}) - S_i(t_n))(S_j(t_{m+1}) - S_j(t_m))] \\
&\quad= \sum_{i,j} E[\xi_i\xi_j](S_i(t_{n+1}) - S_i(t_n))(S_j(t_{m+1}) - S_j(t_m)) \\
&\quad= \sum_i (S_i(t_{n+1}) - S_i(t_n))(S_i(t_{m+1}) - S_i(t_m)) \\
&\overset{\{Parseval\}}{=} \int_0^T 1_{(t_n,t_{n+1})}(s)1_{(t_m,t_{m+1})}(s)ds = \begin{cases} 0 & \text{if } m \ne n, \\ t_{n+1} - t_n & \text{if } n = m. \end{cases}
\end{aligned}$$

To prove uniform convergence, the goal is to establish

$$P\left(\sup_{t\in[0,T]}\sum_{i=1}^{\infty}|\xi_i|S_i(t)<\infty\right)=1.$$

Fix a n and a $t\in[0,T]$ then there is only one i, satisfying $2^n\le i<2^{n+1}$, such that $S_i(t)\neq 0$. Denote this i by $i(t,n)$. Let $\chi_n\equiv\sup_{2^n\le i<2^{n+1}}|\xi_i|$, then

$$\begin{aligned}\sup_{t\in[0,T]}\sum_{i=1}^{\infty}|\xi_i|S_i(t) &= \sup_{t\in[0,T]}\sum_{n=0}^{\infty}|\xi_{i(t,n)}|S_{i(t,n)}(t)\\ &\le \sup_{t\in[0,T]}\sum_{n=0}^{\infty}|\xi_{i(t,n)}|T^{-1/2}2^{-(n+2)/2}\\ &\le \sum_{n=0}^{\infty}\chi_n T^{-1/2}2^{-(n+2)/2}.\end{aligned}$$

If

$$\sum_{n=0}^{\infty}\chi_n 2^{-(n+2)/2}=\infty \tag{2.13}$$

on a set with positive probability, then $\chi_n>n$ for infinitely many n, with positive probability, and consequently

$$\infty=E[\sum_{n=0}^{\infty}1_{\{\chi_n>n\}}]=\sum_{n=0}^{\infty}P(\chi_n>n), \tag{2.14}$$

but

$$P(\chi_n>n)\le P(\cup_{i=2^n}^{2^{n+1}}\{|\xi_i|>n\})\le 2^nP(|\xi_0|>n)\le C\;2^ne^{-n^2/4},$$

so that $\sum_{n=0}^{\infty}P(\chi_n>n)<\infty$, which contradicts (2.14) and (2.13). Therefore $P(\sup_{t\in[0,T]}\sum_{i=1}^{\infty}|\xi_i|S_i(t)<\infty)=1$, which proves the uniform convergence. $\square$

Exercise 2.19 [Extension to multidimensional Itô integrals] The multidimensional Wiener process W in $\mathbb{R}^l$ is defined by $W(t)\equiv(W^1(t),\dots,W^l(t))$, where W^i, $i=1,\dots,l$ are independent one-dimensional Wiener processes. Show that

$$I_{\Delta t}\equiv\sum_{n=0}^{N-1}\sum_{i=1}^{l}f_i(t_n,\cdot)\Delta W_n^i$$

form a Cauchy sequence with $E[(I_{\Delta t_1} - I_{\Delta t_2})^2] = \mathcal{O}(\Delta t_{max})$, as in Theorem 2.10, provided $f : [0,T] \times \Omega \to \mathbb{R}^l$ is adapted and (2.9) holds.

Remark 2.20 A larger class of Itô integrable functions are the functions in the Hilbert space

$$V = \left\{ f : [0,T] \times \Omega \to \mathbb{R}^l : \text{f is adapted and } \int_0^T E[|f(t)|^2]dt < \infty \right\}$$

with the inner product $\int_0^T E[f(t) \cdot g(t)]dt$. This follows from the fact that every function in V can be approximated by adapted functions f_h that satisfy (2.9), for some constant C depending on h, so that $\int_0^T E[|f(t,\cdot) - f_h(t,\cdot)|^2]dt \le h$ as $h \to 0$. However, in contrast to Itô integration of the functions that satisfy (2.9), an approximation of the Itô integrals of $f \in V$ does not in general give a convergence rate, but only convergence.

Exercise 2.21 Read Example 2.18 and show that the Haar-functions can be used to approximate stochastic integrals $\int_0^T f(t)dW(t) \simeq \sum_{i=0}^m \xi_i f_i$, for given deterministic functions f with $f_i = \int_0^T f(s)H_i(s)ds$. In what sense does $dW(s) = \sum_{i=0}^\infty \xi_i H_i ds$ hold?

Exercise 2.22 Give an interpretation of the approximation (2.12) in terms of Brownian bridges, cf. [KS].

Chapter 3

Stochastic Differential Equations

This chapter extends the work on stochastic integrals, in the last chapter, and constructs approximations of stochastic differential equations with an error estimate. Existence and uniqueness is then provided by the error estimate.

We will denote by C, C' positive constants, not necessarily the same at each occurrence.

3.1 Approximation and Definition of SDE

We will prove convergence of Forward Euler approximations of stochastic differential equations, following the convergence proof for Itô integrals. The proof is divided into four steps, including Grönwall's lemma below. The first step extends the Euler approximation $\bar{X}(t)$ to all $t \in [0, T]$:

Step 1. Consider a grid in the interval $[0, T]$ defined by the set of nodes $\{\bar{t}_n\}_{n=0}^{\bar{N}}$, $\bar{t}_0 = 0, \bar{t}_{\bar{N}} = T$ and define the discrete stochastic process $\bar{X}$ by the forward Euler method

$$\bar{X}(\bar{t}_{n+1}) - \bar{X}(\bar{t}_n) = a(\bar{t}_n, \bar{X}(\bar{t}_n))(\bar{t}_{n+1} - \bar{t}_n) + b(\bar{t}_n, \bar{X}(\bar{t}_n))(W(\bar{t}_{n+1}) - W(\bar{t}_n)), \tag{3.1}$$

for $n = 0, \ldots, \bar{N} - 1$. Now extend $\bar{X}$ continuously, for theoretical purposes only, to all values of t by

$$\bar{X}(t) = \bar{X}(\bar{t}_n) + \int_{\bar{t}_n}^{t} a(\bar{t}_n, \bar{X}(\bar{t}_n))ds + \int_{\bar{t}_n}^{t} b(\bar{t}_n, \bar{X}(\bar{t}_n))dW(s),\ \ \bar{t}_n \le t < \bar{t}_{n+1}. \tag{3.2}$$

In other words, the process $\bar{X} : [0,T] \times \Omega \to \mathbb{R}$ satisfies the stochastic differential equation

$$d\bar{X}(s) = \bar{a}(s, \bar{X})ds + \bar{b}(s, \bar{X})dW(s),\ \ \bar{t}_n \le s < \bar{t}_{n+1}, \tag{3.3}$$

where $\bar{a}(s, \bar{X}) \equiv a(\bar{t}_n, \bar{X}(\bar{t}_n)),\ \bar{b}(s, \bar{X}) \equiv b(\bar{t}_n, \bar{X}(\bar{t}_n))$, for $\bar{t}_n \le s < \bar{t}_{n+1}$, and the nodal values of the process $\bar{X}$ is defined by the Euler method (3.1).

Theorem 3.1 *Let $\bar{X}$ and $\bar{\bar{X}}$ be forward Euler approximations of the stochastic process $X : [0,T] \times \Omega \to \mathbb{R}$, satisfying the stochastic differential equation*

$$dX(t) = a(t, X(t))dt + b(t, X(t))dW(t),\ \ 0 \le t < T, \tag{3.4}$$

with time steps

$$\{\bar{t}_n\}_{n=0}^{\bar{N}},\ \ \bar{t}_0 = 0, \bar{t}_{\bar{N}} = T,$$
$$\{\bar{\bar{t}}_m\}_{m=0}^{\bar{\bar{N}}}\ \ \bar{\bar{t}}_0 = 0, \bar{\bar{t}}_{\bar{\bar{N}}} = T,$$

respectively, and

$$\Delta t_{max} = \max\left[\max_{0 \le n \le \bar{N}-1} \bar{t}_{n+1} - \bar{t}_n,\ \max_{0 \le m \le \bar{\bar{N}}-1} \bar{\bar{t}}_{m+1} - \bar{\bar{t}}_m\right].$$

Suppose that there exists a positive constant C such that the initial data and the given functions $a, b : [0,T] \times \mathbb{R} \to \mathbb{R}$ satisfy

$$E[|\bar{X}(0)|^2 + |\bar{\bar{X}}(0)|^2] \le C, \tag{3.5}$$

$$E[\left(\bar{X}(0) - \bar{\bar{X}}(0)\right)^2] \le C\Delta t_{max}, \tag{3.6}$$

and

$$\begin{aligned} |a(t,x) - a(t,y)| &< C|x - y|, \\ |b(t,x) - b(t,y)| &< C|x - y|, \end{aligned} \tag{3.7}$$

$$|a(t,x)-a(s,x)|+|b(t,x)-b(s,x)|\leq C(1+|x|)\sqrt{|t-s|}. \tag{3.8}$$

Then there is a constant K such that

$$\max\left\{E[\bar{X}^2(t,\cdot)],E[\bar{\bar{X}}^2(t,\cdot)]\right\}\leq KT,\ t<T, \tag{3.9}$$

and

$$E\left[\left(\bar{X}(t,\cdot)-\bar{\bar{X}}(t,\cdot)\right)^2\right]\leq K\Delta t_{max},\ t<T. \tag{3.10}$$

The basic idea for the extension of the convergence for Itô integrals to stochastic differntial equations is

Lemma 3.2 (Grönwall) *Assume that there exist positive constants A and K such that the function $f:\mathbb{R}\to\mathbb{R}$ satisfies*

$$f(t)\leq K\int_0^t f(s)ds+A. \tag{3.11}$$

Then

$$f(t)\leq Ae^{Kt}.$$

Proof. Let $I(t)\equiv\int_0^t f(s)ds$. Then by (3.11)

$$\frac{dI}{dt}\leq KI+A,$$

and multiplying by e^{-Kt} we arrive at

$$\frac{d}{dt}(Ie^{-Kt})\leq Ae^{-Kt}.$$

After integrating, and using $I(0)=0$, we obtain $I\leq A\frac{(e^{Kt}-1)}{K}$. Substituting the last result in (3.11) concludes the proof. □

Proof of the Theorem. To prove (3.10), assume first that (3.9) holds. The proof is divided into the following steps:

(1) Representation of $\bar{X}$ as a process in continuous time: Step 1.

(2) Use the assumptions (3.7) and (3.8).

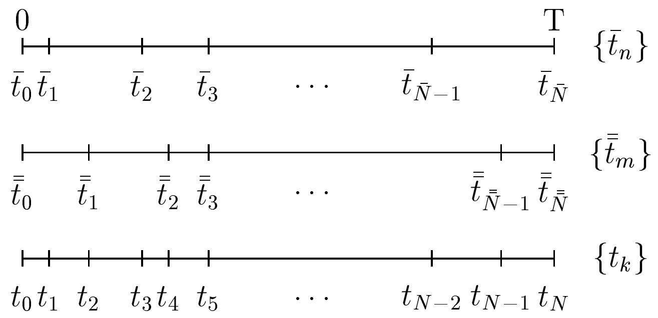


Figure 3.1: Mesh points used in the proof.

(3) Use the property (3) from Theorem 2.15.

(4) Apply Grönwall's lemma.

Step 2. Consider another forward Euler discretization $\bar{\bar{X}}$, defined on a grid with nodes $\left\{\bar{\bar{t}}_m\right\}_{m=0}^{\bar{\bar{N}}}$, and subtract the two solutions to arrive at

$$\bar{X}(s)-\bar{\bar{X}}(s) \overset{(3.3)}{=} \bar{X}(0)-\bar{\bar{X}}(0)+\int_0^s \underbrace{(\bar{a}-\bar{\bar{a}})(t)}_{\equiv\Delta a(t)}\,dt+\int_0^s \underbrace{(\bar{b}-\bar{\bar{b}})(t)}_{\equiv\Delta b(t)}\,dW(t). \quad (3.12)$$

The definition of the discretized solutions implies that

$$\begin{aligned}
\Delta a(t) &= (\bar{a}-\bar{\bar{a}})(t) = a(\bar{t}_n, \bar{X}(\bar{t}_n)) - a(\bar{\bar{t}}_m, \bar{\bar{X}}(\bar{\bar{t}}_m)) = \\
&= \underbrace{a(\bar{t}_n, \bar{X}(\bar{t}_n)) - a(t, \bar{X}(t))}_{=(I)} \\
&\quad + \underbrace{a(t, \bar{X}(t)) - a(t, \bar{\bar{X}}(t))}_{=(II)} \\
&\quad + \underbrace{a(t, \bar{\bar{X}}(t)) - a(\bar{\bar{t}}_m, \bar{\bar{X}}(\bar{\bar{t}}_m))}_{=(III)}
\end{aligned}$$

where $t \in [\bar{\bar{t}}_m, \bar{\bar{t}}_{m+1}) \cap [\bar{t}_n, \bar{t}_{n+1})$, as shown in Figure 3.1. The assumptions (3.7) and (3.8) show that

$$
\begin{aligned}
|(I)| &\le |a(\bar{t}_n, \bar{X}(\bar{t}_n)) - a(t, \bar{X}(\bar{t}_n))| + |a(t, \bar{X}(\bar{t}_n)) - a(t, \bar{X}(t))| \\
&\le C|\bar{X}(\bar{t}_n) - \bar{X}(t)| + C(1 + |\bar{X}(\bar{t}_n)|)|t - \bar{t}_n|^{1/2}.
\end{aligned} \tag{3.13}
$$

Note that (3.7) and (3.8) imply

$$
|a(t,x)| + |b(t,x)| \le C(1 + |x|). \tag{3.14}
$$

Therefore

$$
\begin{aligned}
|\bar{X}(\bar{t}_n) - \bar{X}(t)| &\overset{(3.3)}{=} |a(\bar{t}_n, \bar{X}(\bar{t}_n))(t - \bar{t}_n) + b(\bar{t}_n, \bar{X}(\bar{t}_n))(W(t) - W(\bar{t}_n))| \\
&\overset{(3.14)}{\le} C(1 + |\bar{X}(\bar{t}_n)|)((t - \bar{t}_n) + |W(t) - W(\bar{t}_n)|).
\end{aligned} \tag{3.15}
$$

The combination of (3.13) and (3.15) shows

$$
|(I)| \le C(1 + |\bar{X}(\bar{t}_n)|) \left(|W(t) - W(\bar{t}_n)| + |t - \bar{t}_n|^{1/2}\right)
$$

and in a similar way,

$$
|(III)| \le C(1 + |\bar{\bar{X}}(t)|) \left(|W(t) - W(\bar{\bar{t}}_m)| + |t - \bar{\bar{t}}_m|^{1/2}\right),
$$

and by the assumptions (3.7)

$$
|(II)| \overset{(3.7)}{\le} C|\bar{X}(t) - \bar{\bar{X}}(t)|.
$$

Therefore, the last three inequalities imply

$$
\begin{aligned}
|\Delta a(t)|^2 &\le (|(I)| + |(II)| + |(III)|)^2 \le C_2 \Big(|\bar{X}(t) - \bar{\bar{X}}(t)|^2 \\
&\quad + (1 + |\bar{X}(\bar{t}_n)|^2)(|t - \bar{t}_n| + |W(t) - W(\bar{t}_n)|^2) \\
&\quad + (1 + |\bar{\bar{X}}(\bar{\bar{t}}_m)|^2)(|t - \bar{\bar{t}}_m| + |W(t) - W(\bar{\bar{t}}_m)|^2)\Big).
\end{aligned} \tag{3.16}
$$

Recall that $\max(t - \bar{t}_n, t - \bar{\bar{t}}_m) \le \Delta t_{max}$, and

$$
E[(W(t) - W(s))^2] = t - s, \; s < t,
$$

so that the expected value of (3.16) and the assumption (3.9) yield

$$
\begin{aligned}
E[|\Delta a(t)|^2] &\leq C\left(E[|\bar{X}(t)-\bar{\bar{X}}(t)|^2]+(1+E[|\bar{X}(\bar{t}_n)|^2]+E[|\bar{\bar{X}}(\bar{\bar{t}}_m)|^2])\Delta t_{max}\right)\\
&\overset{(3.9)}{\leq} C\left(E[|\bar{X}(t)-\bar{\bar{X}}(t)|^2]+\Delta t_{max}\right). \qquad (3.17)
\end{aligned}
$$

Similarly, we have

$$
E[|\Delta b(t)|^2] \leq C\left(E[|\bar{X}(t)-\bar{\bar{X}}(t)|^2]+\Delta t_{max}\right). \qquad (3.18)
$$

Step 3. Define a refined grid $\{t_h\}_{h=0}^N$ by the union

$$
\{t_h\} \equiv \{\bar{t}_n\} \cup \{\bar{\bar{t}}_m\}.
$$

Observe that both the functions $\Delta a(t)$ and $\Delta b(t)$ are adapted and piecewise constant on the refined grid. The error representation (3.12) and (3) of Theorem 2.15 imply

$$
\begin{aligned}
E[|\bar{X}(s)-\bar{\bar{X}}(s)|^2] &\leq E\left[\left(\bar{X}(0)-\bar{\bar{X}}(0)+\int_0^s \Delta a(t)dt+\int_0^s \Delta b(t)dW(t)\right)^2\right]\\
&\leq 3E[|\bar{X}(0)-\bar{\bar{X}}(0)|^2]\\
&+ 3E\left[\left(\int_0^s \Delta a(t)dt\right)^2\right]+3E\left[\left(\int_0^s \Delta b(t)dW(t)\right)^2\right]\\
&\overset{(3.6)}{\leq} 3(C\Delta t_{max}+s\int_0^s E[(\Delta a(t))^2]dt+\int_0^s E[(\Delta b(t))^2]dt).
\end{aligned}
\qquad (3.19)
$$

Inequalities (3.17-3.19) combine to

$$
E[|\bar{X}(s)-\bar{\bar{X}}(s)|^2] \overset{(3.17-3.19)}{\leq} C(\int_0^s E[|\bar{X}(t)-\bar{\bar{X}}(t)|^2]dt+\Delta t_{max}). \qquad (3.20)
$$

Step 4. Finally, Grönwall's Lemma 3.2 applied to (3.20) implies

$$
E[|\bar{X}(t)-\bar{\bar{X}}(t)|^2] \leq \Delta t_{max} C e^{Ct},
$$

which finishes the proof. □

Exercise 3.3 Prove (3.9). Hint: Follow Steps 1-4 and use (3.5) .

□

Corollary 3.4 *The previous theorem yields a convergence result also in the* L^2 *norm* $\|X\|^2 = \int_0^T E[X(t)^2]dt$. *The order of this convergence is* $1/2$, *i.e.* $\|\bar{X} - \bar{\bar{X}}\| = \mathcal{O}(\sqrt{\Delta t_{max}})$.

Remark 3.5 [Strong and weak convergence] Depending on the application, our interest will be focused either on strong convergence

$$\|X(T) - \bar{X}(T)\|_{L^2[\Omega]} = \sqrt{E[(X(T) - \bar{X}(T))^2]} = \mathcal{O}(\sqrt{\Delta t}),$$

or on weak convergence $E[g(X(T))] - E[g(\bar{X}(T))]$, for given functions g. The next chapters will show first order convergence of expected values for the Euler method,

$$E[g(X(T)) - g(\bar{X}(T))] = \mathcal{O}(\Delta t),$$

and introduce Monte Carlo methods to approximate expected values $E[g(\bar{X}(T))]$. We will distinguish between strong and weak convergence by $X_n \to X$, denoting the strong convergence $E[|X_n - X|^2] \to 0$ for random variables and $\int_0^T E[|X_n(t) - X(t)|^2]dt \to 0$ for stochastic processes, and by $X_n \rightharpoonup X$, denoting the weak convergence $E[g(X_n)] \to E[g(X)]$ for all bounded continuous functions g.

□

Exercise 3.6 Show that strong convergence, $X_n \to X$, implies weak convergence $X_n \rightharpoonup X$. Show also by an example that weak convergence, $X_n \rightharpoonup X$, does not imply strong convergence, $X_n \to X$. *Hint:* Let $\{X_n\}$ be a sequence of independent identically distributed random variables.

□

Corollary 3.4 shows that successive refinements of the forward Euler approximation forms a Cauchy sequence in the Hilbert space V, defined by

Definition 2.20. The limit $X \in V$, of this Cauchy sequence, satisfies the stochastic equation

$$X(s) = X(0) + \int_0^s a(t, X(t))dt + \int_0^s b(t, X(t))dW(t), \quad 0 < s \le T, \tag{3.21}$$

and it is unique, (why?). Hence, we have constructed existence and uniqueness of solutions of (3.21) by forward Euler approximations. Let X be the solution of (3.21). From now on we use indistinctly also the notation

$$\begin{aligned} dX(t) &= a(t, X(t))dt + b(t, X(t))dW(t), \quad 0 < t \le T \\ X(0) &= X_0. \end{aligned} \tag{3.22}$$

These notes focus on the Euler method to approximate stochastic differential equations (3.22). The following result motivates that there is no method with higher order convergence rate than the Euler method to control the strong error $\int_0^1 E[(X(t) - \bar{X}(t))^2]dt$, since even for the simplest equation $dX = dW$ any linear approximation $\hat{W}$ of W, based on N function evaluations, satisfies

Theorem 3.7 *Let $\hat{W}(t) = f(t, W(t_1), \ldots, W(t_N))$ be any approximation of $W(t)$, which for fixed t is based on any linear function $f(t, \cdot) : \mathbb{R}^N \to \mathbb{R}$, and a partition $0 = t_0 < \ldots < t_N = 1$ of $[0, 1]$, then the strong approximation error is bounded from below by*

$$\left(\int_0^1 E[(W(t) - \hat{W}(t))^2]dt \right)^{1/2} \ge \frac{1}{\sqrt{6N}}, \tag{3.23}$$

which is the same error as for the Euler method based on constant time steps and linear interpolation between the time steps.

Proof. The linearity of $f(t, \cdot)$ implies that

$$\hat{W}(t) \equiv \sum_{i=1}^N \alpha_i(t) \Delta W_i$$

where $\alpha_i : [0, 1] \to \mathbb{R}$, $i = 1, \ldots, N$ are any functions. The idea is to choose the functions $\alpha_i : [0, 1] \to \mathbb{R}$, $i = 1, \ldots, N$ in an optimal way, and see that

Figure 3.2: Optimal choice for weight functions α_i.

the minimum error satisfies (3.23). We have

$$\begin{aligned}
&\int_0^1 E[(W(t) - \hat{W}(t))^2]dt \\
&= \int_0^1 \big(E[W^2(t)] - 2\sum_{i=1}^N \alpha_i(t) E[W(t)\Delta W_i] + \sum_{i,j=1}^N \alpha_i(t)\alpha_j(t) E[\Delta W_i \Delta W_j]\big) dt \\
&= \int_0^1 t dt - 2\int_0^1 \sum_{i=1}^N E[W(t)\Delta W_i]\alpha_i dt + \int_0^1 \sum_{i=1}^N \alpha_i^2(t)\Delta t_i dt
\end{aligned}$$

and in addition

$$E[W(t)\Delta W_i] = \begin{cases} \Delta t_i, & t_{i+1} < t \\ (t - t_i), & t_i < t < t_{i+1} \\ 0, & t < t_i. \end{cases} \tag{3.24}$$

Perturbing the functions α_i, to $\alpha_i + \epsilon\delta_i$, $\epsilon << 1$, around the minimal value of $\int_0^1 E[W(t) - \hat{W}(t)]^2 dt$ gives the following conditions for the optimum choice of α_i, cf. Figure 3.2:

$$-2E[W(t)\Delta W_i] + 2\alpha_i^*(t)\Delta t_i = 0, \; i = 1, \ldots, N.$$

and hence

$$\begin{aligned}
&\min \int_0^1 E[W(t) - \hat{W}(t)]^2 dt = \int_0^1 t dt - \int_0^1 \sum_{i=1}^N \frac{E[W(t)\Delta W_i]^2}{\Delta t_i} dt \\
&\underbrace{=}_{(3.24)} \sum_{n=1}^N (t_n + \Delta t_n/2)\Delta t_n - \sum_{n=1}^N \left(t_n \Delta t_n + \int_{t_n}^{t_{n+1}} \frac{(t - t_n)^2}{\Delta t_n} dt \right) \\
&= \sum_{n=1}^N (\Delta t_n)^2/6 \geq \frac{1}{6N}.
\end{aligned}$$

where Exercise 3.8 is used in the last inequality and proves the lower bound of the approximation error in the theorem. Finally, we note that by (3.24) the optimal $\alpha_i^*(t) = \frac{E[W(t)\Delta W_i]}{\Delta t_i}$ is infact linear interpolation of the Euler method. $\square$

Exercise 3.8 To verify the last inequality in the previous proof, compute

$$\min_{\Delta t} \sum_{n=1}^{N} (\Delta t_n)^2$$
$$\textit{subject to}$$
$$\sum_{n=1}^{N} (\Delta t_n) = 1.$$

$\square$

3.2 Itô's Formula

Recall that using a forward Euler discretization we found the relation

$$\begin{aligned} \int_0^T W(s)dW(s) &= W^2(T)/2 - T/2, \text{ or} \\ W(s)dW(s) &= d(W^2(s)/2) - ds/2, \end{aligned} \tag{3.25}$$

whereas in the deterministic case we have $y(s)dy(s) = d(y^2(s)/2)$. The following useful theorem with Itô 's formula generalizes (3.25) to general functions of solutions to the stochastic differential equations.

Theorem 3.9 *Suppose that the assumptions in Theorem 2.10 hold and that X satisfies the stochastic differential equation*

$$\begin{aligned} dX(s) &= a(s, X(s))ds + b(s, X(s))dW(s), \quad s > 0 \\ X(0) &= X_0, \end{aligned}$$

and let $g : (0, +\infty) \times \mathbb{R} \to \mathbb{R}$ be a given bounded function in $C^2((0,\infty) \times \mathbb{R})$. Then $y(t) \equiv g(t, X(t))$ satisfies the stochastic differential equation

$$\begin{aligned} dy(t) &= \left(\partial_t g(t, X(t)) + a(s, X(s))\partial_x g(t, X(t)) + \frac{b^2(t, X(t))}{2} \partial_{xx} g(t, X(t)) \right) dt \\ &+ b(t, X(t))\partial_x g(t, X(t))dW(t), \end{aligned} \tag{3.26}$$

Proof. We want to prove the Itô formula in the integral sense

$$
\begin{aligned}
&g(\tau, X(\tau)) - g(0, X(0)) \\
&= \int_0^\tau \left(\partial_t g(t, X(t)) + a(s, X(s))\partial_x g(t, X(t)) + \frac{b^2(t, X(t))}{2} \partial_{xx} g(t, X(t)) \right) dt \\
&\quad + \int_0^\tau b(t, X(t))\partial_x g(t, X(t)) dW(t).
\end{aligned}
$$

Let $\bar{X}$ be a forward Euler approximation (3.1) and (3.2) of X, so that

$$
\Delta\bar{X} \equiv \bar{X}(t_n + \Delta t_n) - \bar{X}(t_n) = a(t_n, \bar{X}(t_n))\Delta t_n + b(t_n, \bar{X}(t_n))\Delta W_n. \quad (3.27)
$$

Taylor expansion of g up to second order gives

$$
\begin{aligned}
&g(t_n + \Delta t_n, \bar{X}(t_n + \Delta t_n)) - g(t_n, \bar{X}(t_n)) \\
&= \partial_t g(t_n, \bar{X}(t_n))\Delta t_n + \partial_x g(t_n, \bar{X}(t_n))\Delta\bar{X}(t_n) \\
&\quad + \frac{1}{2}\partial_{tt} g(t_n, \bar{X}(t_n))\Delta t_n^2 + \partial_{tx} g(t_n, \bar{X}(t_n))\Delta t_n \Delta\bar{X}(t_n) \\
&\quad + \frac{1}{2}\partial_{xx} g(t_n, \bar{X}(t_n))(\Delta\bar{X}(t_n))^2 + o(\Delta t_n^2 + |\Delta\bar{X}_n|^2). \quad (3.28)
\end{aligned}
$$

The combination of (3.27) and (3.28) shows

$$
\begin{aligned}
g(t_m, \bar{X}(t_m)) - g(0, \bar{X}(0)) &= \sum_{n=0}^{m-1} \big(g(t_n + \Delta t_n, \bar{X}(t_n + \Delta t_n)) - g(t_n, \bar{X}(t_n)) \big) \\
= &\sum_{n=0}^{m-1} \partial_t g \Delta t_n + \sum_{n=0}^{m-1} (\bar{a}\partial_x g \Delta t_n + \bar{b}\partial_x g \Delta W_n) + \frac{1}{2}\sum_{n=0}^{m-1} (\bar{b})^2 \partial_{xx} g (\Delta W_n)^2 \\
&+ \sum_{n=0}^{m-1} (\bar{b}\partial_{tx} g \Delta t_n \Delta W_n + (\frac{1}{2}\partial_{tt} g + \bar{a}\partial_{tx} g)\Delta t_n^2) \\
&+ \sum_{n=0}^{m-1} o(\Delta t_n^2 + |\Delta\bar{X}(t_n)|^2). \quad (3.29)
\end{aligned}
$$

Let us first show that

$$
\sum_{n=0}^{m-1} (\bar{b})^2 \partial_{xx} g(\bar{X})(\Delta W_n)^2 \to \int_0^t b^2 \partial_{xx} g(X) ds,
$$

as $\Delta t_{max} \to 0$. It is sufficient to establish

$$Y \equiv \frac{1}{2} \sum_{n=0}^{m-1} (\bar{b})^2 \partial_{xx} g((\Delta W_n)^2 - \Delta t_n) \to 0, \tag{3.30}$$

since (3.10) implies $\sum_{n=0}^{m-1} (\bar{b})^2 \partial_{xx} g \Delta t_n \to \int_0^t b^2 \partial_{xx} g ds$. Use the notation $\alpha_i = ((\bar{b})^2 \partial_{xx} g)(t_i, \bar{X}(t_i))$ and independence to obtain

$$\begin{aligned}
E[Y^2] &= \sum_{i,j} E[\alpha_i \alpha_j ((\Delta W_i)^2 - \Delta t_i)((\Delta W_j)^2 - \Delta t_j)] \\
&= 2 \sum_{i>j} E[\alpha_i \alpha_j ((\Delta W_j)^2 - \Delta t_j)((\Delta W_i)^2 - \Delta t_i)] + \sum_i E[\alpha_i^2 ((\Delta W_i)^2 - \Delta t_i)^2] \\
&= 2 \sum_{i>j} E[\alpha_i \alpha_j ((\Delta W_j)^2 - \Delta t_j)] \underbrace{E[((\Delta W_i)^2 - \Delta t_i)]}_{=0} \\
&\quad + \sum_i E[\alpha_i^2] \underbrace{E[((\Delta W_i)^2 - \Delta t_i)^2]}_{=2\Delta t_i^2} \to 0,
\end{aligned}$$

when $\Delta t_{max} \to 0$, therefore (3.30) holds. Similar analysis with the other terms in (3.29) concludes the proof. $\square$

Remark 3.10 The preceding result can be remembered intuitively by a Taylor expansion of g up to second order

$$dg = \partial_t g \ dt + \partial_x g \ dX + \frac{1}{2} \partial_{xx} g \ (dX)^2$$

and the relations: $dtdt = dtdW = dWdt = 0$ and $dWdW = dt$.

Example 3.11 Let $X(t) = W(t)$ and $g(x) = \frac{x^2}{2}$. Then

$$d\left(\frac{W^2(s)}{2}\right) = W(s)dW(s) + 1/2(dW(s))^2 = W(s)dW(s) + ds/2.$$

Exercise 3.12 Let $X(t) = W(t)$ and $g(x) = x^4$. Verify that

$$d(W^4(s)) = 6W^2(s)ds + 4W^3(s)dW(s)$$

and

$$\frac{d}{ds}(E[g(W(s))]) = \frac{d}{ds}(E[(W(s))^4]) = 6s.$$

Apply the last result to compute $E[W^4(t)]$ and $E[(W^2(t) - t)^2]$.

Exercise 3.13 Generalize the previous exercise to deteremine $E[W^{2n}(t)]$.

Example 3.14 We want to compute $\int_0^T tdW(t)$. Take $g(t,x) = tx$, and again $X(t) = W(t)$, so that

$$tW(t) = \int_0^t sdW(s) + \int_0^t W(s)ds$$

and finally $\int_0^t sdW(s) = tW(t) - \int_0^t W(s)ds$.

Exercise 3.15 Consider the stochastic differential equation

$$dX(t) = -a(X(t) - X_\infty)dt + bdW(t),$$

with initial data $X(0) = X_0 \in \mathbb{R}$ and given $a, b \in \mathbb{R}$.

(i) Using that

$$X(t) - X(0) = -a\int_0^t (X(s) - X_\infty)dt + bW(t),$$

take the expected value and find an ordinary differential equation for the function $m(t) \equiv E[X(t)]$.

(ii) Use Itô 's formula to find the differential of $(X(t))^2$ and apply similar ideas as in (i) to compute $Var[X(t)]$.

(iii) Use an integrating factor to derive the exact solution (2.10) in Example 2.17. Compare your results from (i) and (ii) with this exact solution.

Example 3.16 Consider the stochastic differential equation

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

used to model the evolution of stock values. The values of r (interest rate) and σ (volatility) are assumed to be constant. Our objective is to find a closed expression for the solution, often called *geometric Brownian motion*. Let $g(x) = \ln(x)$. Then a direct application of Itô formula shows

$$d\ln(S(t)) = dS(t)/S(t) - 1/2\left(\frac{\sigma^2 S^2(t)}{S^2(t)}\right)dt = rdt - \frac{\sigma^2}{2}dt + \sigma dW(t),$$

so that

$$\ln\left(\frac{S(T)}{S(0)}\right) = rT - \frac{T\sigma^2}{2} + \sigma W(T)$$

and consequently

$$S(T) = e^{(r-\frac{\sigma^2}{2})T+\sigma W(T)}S(0).$$

Exercise 3.17 Suppose that we want to simulate $S(t)$, defined in the previous example by means of the forward Euler method, i.e.

$$S_{n+1} = (1 + r\Delta t_n + \sigma \Delta W_n)S_n, \ n = 0, \ldots, N$$

As with the exact solution $S(t)$, we would like to have S_n positive. Then we could choose the time step Δt_n to reduce the probability of hitting zero

$$P(S_{n+1} < 0 | S_n = s) < \epsilon << 1. \tag{3.31}$$

Motivate a choice for ϵ and find then the largest Δt_n satisfying (3.31).

Remark 3.18 The Wiener process has unbounded variation i.e.

$$E\left[\int_0^T |dW(s)|\right] = +\infty.$$

This is the reason why the forward and backward Euler methods give different results. We have for a uniform mesh $\Delta t = T/N$

$$\begin{aligned} E[\sum_{i=0}^{N-1} |\Delta W_i|] &= \sum_{i=0}^{N-1} E[|\Delta W_i|] = \sum_{i=0}^{N-1} \sqrt{\frac{2\Delta t_i}{\pi}} \\ &= \sqrt{\frac{2T}{\pi}} \sum_{i=0}^{N-1} \sqrt{1/N} = \sqrt{\frac{2NT}{\pi}} \to \infty, \quad \text{as } N \to \infty. \end{aligned}$$

3.3 Stratonovich Integrals

Recall from Chapter 2 that Itô integrals are constructed via forward Euler discretizations and Stratonovich integrals via the trapezoidal method, see Exercise 3.19. Our goal here is to express a Stratonovich integral

$$\int_0^T g(t, X(t)) \circ dW(t)$$

in terms of an Itô integral. Assume then that $X(t)$ satisfies the Itô differential equation

$$dX(t) = a(t, X(t))dt + b(t, X(t))dW(t).$$

Then the relation reads

$$\begin{aligned}\int_0^T g(t,X(t))\circ dW(t) &= \int_0^T g(t,X(t))dW(t)\\ &+ \frac{1}{2}\int_0^T \partial_x g(t,X(t))b(t,X(t))dt. \end{aligned}\qquad (3.32)$$

Therefore, Stratonovich integrals satisfy

$$dg(t,X(t)) = \partial_t g(t,X(t))dt + \partial_x g(t,X(t))\circ dX(t), \qquad (3.33)$$

just like in the usual calculus.

Exercise 3.19 Use that Stratonovich integrals $g(t,X(t))\circ dW(t)$ are defined by limits of the trapezoidal method to verify (3.32), cf. Remark 2.9.

Exercise 3.20 Verify the relation (3.33), and use this to show that $dS(t) = rS(t)dt + \sigma S(t)\circ dW(t)$ implies $S(t) = e^{rt+\sigma W(t)}S(0)$.

Remark 3.21 [Stratonovich as limit of piecewise linear interpolations] Let $R^N(t) \equiv W(t_n) + \frac{W(t_{n+1})-W(t_n)}{t_{n+1}-t_n}(t-t_n),\ t\in(t_n,t_{n+1})$ be a piecewise linear interpolation of W on a given grid, and define X^N by $dX^N(t) = a(X^N(t))dt + b(X^N(t))dR^N(t)$. Then $X^N \to X$ in L^2, where X is the solution of the Stratonovich stochastic differential equation

$$dX(t) = a(X(t))dt + b(X(t))\circ dW(t).$$

In the special case when $a(x) = rx$ and $b(x) = \sigma x$ this follows from

$$\frac{d}{dt}(\ln(X^N(t))) = rdt + \sigma dR^N,$$

so that

$$X^N(t) = e^{rt+\sigma R^N(t)}X(0).$$

The limit $N\to\infty$ implies $X^N(t)\to X(t) = e^{rt+\sigma W(t)}X(0)$, as in Exercise 3.20.

3.4 Systems of SDE

Let $W_1, W_2, \ldots, W_l$ be scalar independent Wiener processes. Consider the l-dimensional Wiener process $W = (W_1, W_2, \ldots, W_l)$ and $X : [0,T] \times \Omega \to \mathbb{R}^d$ satisfying for given drift $a : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ and diffusion $b : [0,T] \times \mathbb{R}^d \to \mathbb{R}^{d\times l}$ the Itô stochastic differential equation

$$dX_i(t) = a_i(t, X(t))dt + b_{ij}(t, X(t))dW_j(t), \; for \; i = 1 \ldots d. \tag{3.34}$$

Here and below we use of the summation convention

$$\alpha_j \beta_j \equiv \sum_j \alpha_j \beta_j,$$

i.e., if the same summation index appears twice in a term, the term denotes the sum over the range of this index. Theorem 3.9 can be directly generalized to the system (3.34).

Theorem 3.22 (Itô 's formula for systems) *Let*

$$dX_i(t) = a_i(t, X(t))dt + b_{ij}(t, X(t))dW_j(t), \; for \; i = 1 \ldots d,$$

and consider a smooth and bounded function $g : \mathbb{R}^+ \times \mathbb{R}^d \to \mathbb{R}$. *Then*

$$\begin{aligned} dg(t, X(t)) \quad = \quad & \Big\{ \partial_t g(t, X(t)) + \partial_{x_i} g(t, X(t)) a_i(t, X(t)) \\ & + \frac{1}{2} b_{ik}(t, X(t)) \partial_{x_i x_j} g(t, X(t)) b_{jk}(t, X(t)) \Big\} dt \\ & + \partial_{x_i} g(t, X(t)) b_{ij}(t, X(t)) dW_j(t), \end{aligned}$$

or in matrix vector notation

$$\begin{aligned} dg(t, X(t)) \quad = \quad & \Big\{ \partial_t g(t, X(t)) + \nabla_x g(t, X(t)) \; a(t, X(t)) \\ & + \frac{1}{2} \mathrm{trace} \left(\mathrm{b(t, X(t)) b^T (t, X(t)) \nabla_x^2 g(t, X(t))} \right) \Big\} dt \\ & + \nabla_x g(t, X(t)) \; b(t, X(t)) dW(t). \end{aligned}$$

□

Remark 3.23 The formal rules to remember Theorem 3.22 are Taylor expansion to second order and

$$\begin{aligned} dW_j dt &= dtdt = 0 \\ dW_i dW_j &= \delta_{ij} dt = \begin{cases} dt & if\ i = j, \\ 0 & otherwise. \end{cases} \end{aligned} \tag{3.35}$$

Exercise 3.24 Verify Remark 3.23.

Chapter 4

The Feynman-Kăc Formula and the Black-Scholes Equation

4.1 The Feynman-Kăc Formula

Theorem 4.1 *Suppose that a, b and g are smooth and bounded functions. Let X be the solution of the stochastic differential equation,* $dX(t) = a(t, X(t))dt + b(t, X(t))dW(t)$ *and let* $u(x,t) = E[g(X(T))|X(t) = x]$. *Then u is the solution of the Kolmogorov backward equation*

$$\begin{aligned} L^*u &\equiv u_t + au_x + \frac{1}{2}b^2u_{xx} = 0, \quad t < T \\ u(x,T) &= g(x). \end{aligned} \tag{4.1}$$

Proof. Define $\hat{u}$ to be the solution of (4.1), i.e. $L^*\hat{u} = 0$, $\hat{u}(\cdot, T) = g(\cdot)$. We want to verify that $\hat{u}$ is the expected value $E[g(X(T))|\ X(t) = x]$. The Itô formula applied to $\hat{u}(X(t), t)$ shows

$$\begin{aligned} d\hat{u}(X(t), t) &= \left(\hat{u}_t + a\hat{u}_x + \frac{1}{2}b^2\hat{u}_{xx}\right)dt + b\hat{u}_x dW \\ &= L^*\hat{u}dt + b\hat{u}_x dW. \end{aligned}$$

Integrate this from t to T and use $L^*\hat{u} = 0$ to obtain

$$\begin{aligned} \hat{u}(X(T), T) - \hat{u}(X(t), t) &= g(X(T)) - \hat{u}(X(t), t) \\ &= \int_t^T b\hat{u}_x dW(s). \end{aligned}$$

Take the expectation and use that the expected value of the Itô integral is zero,

$$\begin{aligned} E[g(X(T))|X(t)=x] - \hat{u}(x,t) &= E[\int_t^T b(s,X(s))\hat{u}_x(X(s),s)dW(s)|X(t)=x] \\ &= 0. \end{aligned}$$

Therefore

$$\hat{u}(x,t) = E[g(X(T))|X(t)=x],$$

which proves the theorem since the solution of Equation (4.1) is unique. □

Exercise 4.2 [Maximum Principle] Let the function u satisfy

$$\begin{aligned} u_t + au_x + \frac{1}{2}b^2 u_{xx} &= 0, \quad t < T \\ u(x,T) &= g(x). \end{aligned}$$

Prove that u satisfies the maximum principle

$$\max_{0<t<T,\ x\in\mathbb{R}} u(t,x) \le \max_{x\in\mathbb{R}} g(x).$$

4.2 Black-Scholes Equation

Example 4.3 Let $f(t,S(t))$ be the price of a European put option where $S(t)$ is the price of a stock satisfying the stochastic differential equation $dS = \mu S dt + \sigma S dW$, where the volatility σ and the drift μ are constants. Assume also the existence of a risk free paper, B, which follows $dB = rBdt$, where r, the risk free rent is a constant. Find the partial differential equation of the price, $f(t,S(t))$, of an option.

Solution. Consider the portfolio $I = -f + \alpha\ S + \beta B$ for $\alpha(t), \beta(t) \in \mathbb{R}$. Then the Itô formula and self financing, i.e. $dI = -df + \alpha dS + \beta dB$, imply

$$\begin{aligned} dI &= -df + \alpha dS + \beta dB \\ &= -(f_t + \mu S f_S + \frac{1}{2}\sigma^2 S^2 f_{SS})dt - f_S \sigma S dW + \alpha(\mu S dt + \sigma S dW) + \beta r B dt \\ &= \left(-(f_t + \mu S f_S + \frac{1}{2}\sigma^2 S^2 f_{SS}) + (\alpha\mu S + \beta r B)\right) dt + (-f_S + \alpha)\sigma S dW. \end{aligned}$$

Now choose α such that the portfolio I becomes riskless, i.e. $\alpha = f_S$, so that

$$\begin{aligned} dI &= \Big(-(f_t + \mu S f_S + \frac{1}{2}\sigma^2 S^2 f_{SS}) + (f_S \mu S + \beta r B)\Big)\, dt \\ &= \Big(-(f_t + \frac{1}{2}\sigma^2 S^2 f_{SS}) + \beta r B\Big)\, dt. \end{aligned} \tag{4.2}$$

Assume also that the existence of an arbitrage opportunity is precluded, i.e. $dI = rIdt$, where r is the interest rate for riskless investments, to obtain

$$\begin{aligned} dI &= r(-f + \alpha S + \beta B)dt \\ &= r(-f + f_S S + \beta B)dt. \end{aligned} \tag{4.3}$$

Equation (4.2) and (4.3) show that

$$f_t + rsf_s + \frac{1}{2}\sigma^2 s^2 f_{ss} = rf, \quad t < T, \tag{4.4}$$

and finally at the maturity time T the contract value is given by definition, e.g. a standard European put option satisfies for a given exercise price K

$$f(T, s) = \max(K - s, 0).$$

The deterministic partial differential equation (4.4) is called the Black-Scholes equation. The existence of adapted β is shown in the exercise below. $\square$

Exercise 4.4 Replicating portfolio. It is said that the self financing portfolio, $\alpha S + \beta B$, replicates the option f. Show that there exists an adapted stochastic process $\beta(t)$, with $\beta(0) = -f_S(0, S(0))$, satisfying self financing, $d(\alpha S + \beta B) = \alpha dS + \beta dB$, with $\alpha = f_S$.

Exercise 4.5 Verify that the corresponding equation (4.4) holds if μ, σ and r are given functions of time and stock price.

Exercise 4.6 Simulation of a replicating portfolio. Assume that the previously described Black-Scholes model holds and consider the case of a bank that has written (sold) a call option on the stock S with the parameters

$$S(0) = S_0 = 760, \ r = 0.06, \ \sigma = 0.65, \ K = S_0.$$

with an exercise date, $T = 1/4$ years. The goal of this exercise is to simulate the replication procedure described in Exercise 4.4, using the exact solution of the Black Scholes call price, computed by the Matlab code

```
% BS call option computation
function y = bsch(S,T,K,r,sigma);

normal = inline('(1+erf(x/sqrt(2)))/2','x');
d1 = (log(S/K)+(r+.5*sigma^2)*T)/sigma/sqrt(T);
d2 = (log(S/K)+(r-.5*sigma^2)*T)/sigma/sqrt(T);
y = S*normal(d1)-K*exp(-r*T)*normal(d2);
```

To this end, choose a number of hedging dates, N, and time steps $\Delta t \equiv T/N$. Assume that $\beta(0) = -f_S(0, S_0)$ and then

- Write a code that computes the $\Delta \equiv \partial f(0, S_0)/\partial S_0$ of a call option.
- Generate a realization for $S(n\Delta t, \omega), n = 0, \ldots, N$.
- Generate the corresponding time discrete realizations for the processes α_n and β_n and the portfolio value, $\alpha_n S_n + \beta_n B_n$.
- Generate the value after settling the contract at time T,

$$\alpha_N S_N + \beta_N B_N - \max(S_N - K, 0).$$

Compute with only one realization, and several values of N, say $N = 10, 20, 40, 80$. What do you observe? How would you proceed if you don't have the exact solution of the Black-Scholes equation?

Theorem 4.7 (Feynman-Kǎc) *Suppose that a, b, g, h and V are bounded smooth functions. Let X be the solution of the stochastic differential equation $dX(t) = a(t, X(t))dt + b(t, X(t))dW(t)$ and let*

$$\begin{aligned} u(x,t) &= E[g(X(T))e^{\int_t^T V(s,X(s))ds}|X(t)=x] \\ &+ E[-\int_t^T h(s,X(s))e^{\int_t^s V(\tau,X(\tau))d\tau}ds|X(t)=x]. \end{aligned}$$

Then u is the solution of the partial differential equation

$$\begin{aligned} L_V^* u \equiv u_t \; + \; a u_x \; + \; \frac{1}{2} b^2 u_{xx} \; + \; Vu \;&=\; h, \quad t < T \\ u(x,T) \;&=\; g(x). \end{aligned} \tag{4.5}$$

Proof. Define $\hat{u}$ to be the solution of the equation (4.5), i.e. $L_V^* \hat{u} = h$ and let $G(s) \;\equiv\; e^{\int_t^s V(\tau, X(\tau)) d\tau}$. We want to verify that $\hat{u}$ is the claimed expected value. We have by Itô 's formula, with $L^* \hat{u} = \hat{u}_t + a\hat{u}_x + \frac{1}{2} b^2 \hat{u}_{xx}$,

$$\begin{aligned} d(\hat{u}(s, X(s)) e^{\int_t^s V(\tau, X(\tau)) d\tau}) &= d(\hat{u}(s, X(s)) G) \\ &= G d\hat{u} \; + \; \hat{u} dG \\ &= G(L^* \hat{u} dt \; + \; b\hat{u}_x dW) \; + \; \hat{u} V G dt, \end{aligned}$$

Integrate both sides from t to T, take the expected value and use $L^* \hat{u} = L_V^* \hat{u} - V\hat{u} = h - V\hat{u}$ to obtain

$$\begin{aligned} E[g(X(t))G(T) \quad | \quad & X(t) = x] - \hat{u}(x,t) \\ &= E[\int_t^T GL^* \hat{u} \; ds] \; + \; E[\int_t^T bG\hat{u}_x \; dW] \; + \; E[\int_t^T \hat{u} VG \; ds] \\ &= E[\int_t^T hG \; ds] \; - \; E[\int_t^T \hat{u} VG \; ds] \; + \; E[\int_t^T \hat{u} VG \; ds] \\ &= E[\int_t^T hG \; ds | X(t) = x]. \end{aligned}$$

Therefore

$$\hat{u}(x,t) \;=\; E[g(X(T))G(T) | X(t) = x] \; - \; E[\int_t^T hG \; ds | X(t) = x].$$

$\square$

Remark 4.8 Compare Black-Scholes equation (4.4) with Equation (4.5): then u corresponds to f, X to $\tilde{S}$, $a(t,x) = rx$, $b(t,x) = \sigma x$, $V = -r$ and $h = 0$. Using the Feynman-Kac formula, we obtain
$f(t, \tilde{S}(t)) = E[e^{-r(T-t)} \max(K - \tilde{S}(T), 0)]$, with $d\tilde{S} = r\tilde{S} dt + \sigma \tilde{S} dW$, which establishes the important relation between approximation based on the Monte Carlo method and partial differential equations discussed in Chapter 1.

Corollary 4.9 *Let* $u(x,t) = E[g(X(T))|X(t) = x] = \int_{\mathbb{R}} g(y)P(y,T;x,t)\,dy$. *Then the density function,* P*, solves the Kolmogorov forward equation, also called the Focker-Planck equation,*

$$\begin{aligned} LP &\equiv -\partial_s P - \partial_y(aP) + \partial_y^2\left(\frac{b^2P}{2}\right) = 0, \quad s > t \\ P(y,t;x,t) &= \delta(x-y), \end{aligned}$$

where δ *is the Dirac-delta measure concentrated at zero.*

Proof. Assume $L\hat{P} = 0$, $\hat{P}(y,t;x,t) = \delta(x-y)$. The Feynman-Kǎc formula implies $L^*u = 0$, so that integration by part shows

$$\begin{aligned} 0 &= \int_t^T \int_{\mathbb{R}} L^*_{y,s}u(y,s)\hat{P}(y,s;x,t)\,dyds \\ &= \left[\int_{\mathbb{R}} u(y,s)\hat{P}(y,s;x,t)\,dy\right]_{s=t}^{s=T} + \int_t^T \int_{\mathbb{R}} u(y,s)L_{y,s}\hat{P}(y,s;x,t)\,dyds \\ &= \left[\int_{\mathbb{R}} u(y,s)\hat{P}(y,s;x,t)\,dy\right]_{s=t}^{s=T}. \end{aligned}$$

Consequently,

$$\begin{aligned} u(x,t) &= \int_{\mathbb{R}} g(y)\hat{P}(y,T;x,t)\,dy \\ &= E[g(X(T))|X(t) = x], \end{aligned}$$

for all functions g. Therefore $\hat{P}$ is the density function P. Hence P solves $LP = 0$. $\square$

Exercise 4.10 [Limit probability distribution] Consider the Ornstein-Uhlenbeck process defined by

$$\begin{aligned} dX(s) &= -(m-x)dt + \sqrt{2}dW(s), \\ X(0) &= x_0. \end{aligned}$$

Verify by means of the Focker-Plank equation that there exist a limit distribution for X(s), when $s \to \infty$.

Exercise 4.11 Assume that $S(t)$ is the price of a single stock. Derive a Monte-Carlo and a PDE method to determine the price of a contingent claim with the contract $\int_0^T h(t, S(t))\, dt$, for a given function h, replacing the usual contract $\max(S(T) - K, 0)$ for European call options.

Exercise 4.12 Derive the Black-Scholes equation for a general system of stocks $S(t) \in \mathbb{R}^d$ solving

$$dS_i = a_i(t, S(t))dt + \sum_{j=1}^{d} b_{ij}(t, S(t))dW_j(t)$$

and a rainbow option with the contract $f(T, S(T)) = g(S(T))$ for a given function $g : \mathbb{R}^d \to \mathbb{R}$, for example

$$g(S) = \max\left(\frac{1}{d}\sum_{i=1}^{d} S_i - K, 0\right).$$

Chapter 5

The Monte-Carlo Method

5.1 Statistical Error

This chapter gives the basic understanding of simulation of expected values $E[g(X(T)]$ for a solution, X, of a given stochastic differential equation with a given function g. In general the approximation error has the two parts of statistical error and time discretization error, which are analyzed in the next sections. The estimation of statistical error is based on the Central Limit Theorem. The error estimate for the time discretization error of the Euler method is directly related to the proof of Feyman-Kăc's theorem with an additional residual term measuring the accuracy of the approximation, which turns out to be first order in contrast to the half order accuracy for strong approximation.

Consider the stochastic differential equation

$$dX(t) = a(t, X(t))dt + b(t, X(t))dW(t)$$

on $t_0 \le t \le T$, how can one compute the value $E[g(X(T))]$? The Monte-Carlo method is based on the approximation

$$E[g(X(T))] \simeq \sum_{j=1}^{N} \frac{g(\overline{X}(T;\omega_j))}{N},$$

where $\overline{X}$ is an approximation of X, e.g. the Euler method. The error in the

Monte-Carlo method is

$$
\begin{aligned}
&E[g(X(T))] - \sum_{j=1}^{N} \frac{g(\overline{X}(T;\omega_j))}{N} \\
&= E[g(X(T)) - g(\overline{X}(T))] - \sum_{j=1}^{N} \frac{g(\overline{X}(T;\omega_j)) - E[g(\overline{X}(T))]}{N}. \quad (5.1)
\end{aligned}
$$

In the right hand side of the error representation (5.1), the first part is the time discretization error, which we will consider in the next subsection, and the second part is the statistical error, which we study here.

Example 5.1 Compute the integral $I = \int_{[0,1]^d} f(x)dx$ by the Monte Carlo method, where we assume $f(x) : [0,1]^d \to \mathbf{R}$.

Solution. We have

$$
\begin{aligned}
I &= \int_{[0,1]^d} f(x)\ dx \\
&= \int_{[0,1]^d} f(x)p(x)\ dx \quad (\text{ where } p \text{ is the uniform density function}) \\
&= E[f(x)] \qquad\qquad (\text{ where } x \text{ is uniformly distributed in } [0,1]^d) \\
&\simeq \sum_{n=1}^{N} \frac{f(x(\omega_n))}{N} \\
&\equiv I_N,
\end{aligned}
$$

where $\{x(\omega_n)\}$ is sampled uniformly in the cube $[0,1]^d$, by sampling the components $x_i(\omega_n)$ independent and uniformly on the interval $[0,1]$. $\square$

The Central Limit Theorem is the fundamental result to understand the statistical error of Monte Carlo methods.

Theorem 5.2 (The Central Limit Theorem) *Assume* ξ_n, $n = 1,2,3,\ldots$ *are independent, identically distributed (i.i.d) and* $E[\xi_n] = 0$, $E[\xi_n^2] = 1$. *Then*

$$
\sum_{n=1}^{N} \frac{\xi_n}{\sqrt{N}} \rightharpoonup \nu, \tag{5.2}
$$

where ν is $N(0,1)$ and $\rightharpoonup$ denotes convergence of the distributions, also called weak convergence, i.e. the convergence (5.2) means $E[g(\sum_{n=1}^N \xi_n/\sqrt{N})] \to E[g(\nu)]$ for all bounded and continuous functions g.

Proof. Let $f(t) = E[e^{it\xi_n}]$. Then

$$f^{(m)}(t) = E[i^m \xi_n^m e^{it\xi_n}], \tag{5.3}$$

and

$$\begin{aligned}
E[e^{it\sum_{n=1}^N \xi_n/\sqrt{N}}] &= f\left(\frac{t}{\sqrt{N}}\right)^N \\
&= \left(f(0) + \frac{t}{\sqrt{N}}f'(0) + \frac{1}{2}\frac{t^2}{N}f''(0) + o\left(\frac{t^2}{N}\right)\right)^N.
\end{aligned}$$

The representation (5.3) implies

$$\begin{aligned}
f(0) &= E[1] = 1, \\
f'(0) &= iE[\xi_n] = 0, \\
f''(0) &= -E[\xi_n^2] = -1.
\end{aligned}$$

Therefore

$$\begin{aligned}
E[e^{it\sum_{n=1}^N \xi_n/\sqrt{N}}] &= \left(1 - \frac{t^2}{2N} + o\left(\frac{t^2}{N}\right)\right)^N \\
&\to e^{-t^2/2}, \quad \text{as } N \to \infty \\
&= \int_{\mathbb{R}} \frac{e^{itx}e^{-x^2/2}}{\sqrt{2\pi}}dx, \qquad (5.4)
\end{aligned}$$

and we conclude that the Fourier transform (i.e. the characteristic function) of $\sum_{n=1}^N \xi_n/\sqrt{N}$ converges to the right limit of Fourier transform of the standard normal distribution. It is a fact, cf. [D], that convergence of the Fourier transform together with continuity of the limit Fourier transform at 0 implies weak convergence, so that $\sum_{n=1}^N \xi_n/\sqrt{N} \rightharpoonup \nu$, where ν is $N(0,1)$. The exercise below verifies this last conclusion, without reference to other results. □

Exercise 5.3 Show that (5.4) implies

$$E[g(\sum_{n=1}^{N} \xi_n/\sqrt{N})] \to E[g(\nu)] \tag{5.5}$$

for all bounded continuous functions g. Hint: study first smooth and quickly decaying functions g_s, satisying $g_s(x) = \int_{-\infty}^{\infty} e^{-itx}\hat{g}_s(t)dt/(2\pi)$ with the Fourier transform $\hat{g}_s$ of g_s satisfying $\hat{g}_s \in L^1(\mathbb{R})$; show that (5.4) implies

$$E[g_s(\sum_{n=1}^{N} \xi_n/\sqrt{N})] \to E[g_s(\nu)];$$

then use Chebychevs inequality to verify that no mass of $\sum_{n=1}^{N} \xi_n/\sqrt{N}$ escapes to infinity; finally, let $\chi(x)$ be a smooth cut-off function which is one for $|x| \le N$ and zero for $|x| > 2N$ and split the general bounded continuous function g into $g = g_s + g(1-\chi) + (g\chi - g_s)$, where g_s is an arbitrary close approximation to $g\chi$; use the conclusions above to prove (5.5).

Example 5.4 What is the error of $I_N - I$ in Example 5.1?

Solution. Let the error ϵ_N be defined by

$$\begin{aligned}
\epsilon_N &= \sum_{n=1}^{N} \frac{f(x_n)}{N} - \int_{[0,1]^d} f(x)dx \\
&= \sum_{n=1}^{N} \frac{f(x_n) - E[f(x)]}{N}.
\end{aligned}$$

By the Central Limit Theorem, $\sqrt{N}\epsilon_N \rightharpoonup \sigma\nu$, where ν is $N(0,1)$ and

$$\begin{aligned}
\sigma^2 &= \int_{[0,1]^d} f^2(x)dx - \left(\int_{[0,1]^d} f(x)dx\right)^2 \\
&= \int_{[0,1]^d} \left(f(x) - \int_{[0,1]^d} f(x)dx\right)^2 dx.
\end{aligned}$$

In practice, σ^2 is approximated by

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{n=1}^{N}\left(f(x_n) - \sum_{m=1}^{N}\frac{f(x_m)}{N}\right)^2.$$

$\square$

One can generate approximate random numbers, so called pseudo random numbers, by for example the method

$$\xi_{i+1} \equiv a\xi_i + b \mod n$$

where a and n are relative prime and the initial ξ_0 is called the seed, which determines all other ξ_i. For example the combinations $n = 2^{31}$, $a = 2^{16} + 3$ and $b = 0$, or $n = 2^{31} - 1$, $a = 7^5$ and $b = 0$ are used in practise. In Monte Carlo computations, we use the pseudo random numbers $\{x_i\}_{i=1}^N$, where $x_i = \frac{\xi_i}{n} \in [0, 1]$, which for $N \ll 2^{31}$ behave approximately as independent uniformly distributed variables.

Theorem 5.5 *The following Box-Müller method generates two independent normal random variables x_1 and x_2 from two independent uniformly distributed variables y_1 and y_2*

$$\begin{aligned} x_1 &= \sqrt{-2\log(y_2)}\ \cos(2\pi y_1) \\ x_2 &= \sqrt{-2\log(y_2)}\ \sin(2\pi y_1). \end{aligned}$$

Sketch of the Idea. The variables x and y are independent standard normal variables if and only if their joint density function is $e^{-(x^2+y^2)/2}/2\pi$. We have

$$e^{-(x^2+y^2)/2}dxdy = re^{-r^2/2}drd\theta = d(e^{-r^2/2})d\theta$$

using $x = rcos\theta$, $y = rsin\theta$ and $0 \le \theta < 2\pi$, $0 \le r < \infty$. The random variables θ and r can be sampled by taking θ to be uniformly distributed in the interval $[0, 2\pi)$ and $e^{-r^2/2}$ to be uniformly distributed in $(0, 1]$, i.e. $\theta = 2\pi y_1$, and $r = \sqrt{-2log(y_2)}$. $\square$

Example 5.6 Consider the stochastic differential equation $dS = rSdt + \sigma SdW$, in the risk neutral formulation where r is the riskless rate of return and σ is the volatility. Then

$$S_T = S_0\ e^{rT - \frac{\sigma^2}{2}T + \sigma\sqrt{T}\nu}$$

where ν is $N(0,1)$. The values of a call option, f_c, and put option, f_p, are by Remark 4.8

$$f_c = e^{-rT}E[\max(S(T)-K,0)]$$

and

$$f_p = e^{-rT}E[\max(K-S(T),0)].$$

□

Example 5.7 Consider the system of stochastic differential equations,

$$dS_i = rS_i dt + \sum_{j=1}^{M} \sigma_{ij} S_i dW_j, \qquad i = 1, ..., M.$$

Then

$$S_i(T) = S_i(0)\ e^{rT - \sum_{j=1}^{M}\left(\sigma_{ij}\sqrt{T}\nu_j - \frac{\sigma_{ij}^2}{2}T\right)}$$

where ν_j are independent and $N(0,1)$. A rainbow call option, based on $S_{av} = \frac{1}{M}\sum_{i=1}^{M} S_i$, can then be simulated by the Monte Carlo method and

$$f_c = e^{-rT}E[\max(S_{av}(T)-K,0)].$$

□

5.2 Time Discretization Error

Consider the stochastic differential equation

$$dX(t) = a(t,X(t))dt + b(t,X(t))dW(t), \qquad 0 \le t \le T,$$

and let $\overline{X}$ be the forward Euler discretization of X. Then

$$\overline{X}(t_{n+1}) - \overline{X}(t_n) = a(t_n, \overline{X}(t_n))\Delta t_n + b(t_n, \overline{X}(t_n))\Delta W_n, \tag{5.6}$$

where $\Delta t_n = t_{n+1} - t_n$ and $\Delta W_n = W(t_{n+1}) - W(t_n)$ for a given discretization $0 = t_0 < t_1 < ... < t_N = T$. Equation (5.6) can be extended, for theoretical use, to all t by

$$\overline{X}(t) - \overline{X}(t_n) = \int_{t_n}^{t} \bar{a}(s, \overline{X}) ds + \int_{t_n}^{t} \bar{b}(s, \overline{X}) dW(s), \quad t_n \le t < t_{n+1},$$

where, for $t_n \le s < t_{n+1}$,

$$\begin{aligned} \bar{a}(s, \overline{X}) &= a(t_n, \overline{X}(t_n)), \\ \bar{b}(s, \overline{X}) &= b(t_n, \overline{X}(t_n)). \end{aligned} \tag{5.7}$$

Theorem 5.8 *Assume that a, b and g are smooth and decay sufficiently fast as $|x| \to \infty$. Then there holds*

$$E[g(X(T)) - g(\overline{X}(T))] = \mathcal{O}(\max \Delta t).$$

Proof. Let u satisfy the equation

$$L^* u \equiv u_t + a u_x + \frac{b^2}{2} u_{xx} = 0, \qquad t < T \tag{5.8}$$

$$u(x, T) = g(x). \tag{5.9}$$

The Feynman-Kăc formula shows

$$u(x, t) = E[g(X(T)) | X(t) = x]$$

and in particular

$$u(0, X(0)) = E[g(X(T))]. \tag{5.10}$$

Then by the Itô formula,

$$\begin{aligned} du(t, \overline{X}(t)) &= \left(u_t + \bar{a} u_x + \frac{\bar{b}^2}{2} u_{xx} \right)(t, \overline{X}(t)) dt + \bar{b} u_x(t, \overline{X}(t)) dW \\ &\overset{(5.8)}{=} \left(-a u_x - \frac{b^2}{2} u_{xx} + \bar{a} u_x + \frac{\bar{b}^2}{2} u_{xx} \right)(t, \overline{X}(t)) dt + \bar{b} u_x(t, \overline{X}(t)) dW \\ &= \left\{ (\bar{a} - a) u_x(t, \overline{X}(t)) + \left(\frac{\bar{b}^2}{2} - \frac{b^2}{2} \right) u_{xx}(t, \overline{X}(t)) \right\} dt \\ &+ \bar{b}(t, \overline{X}) u_x(t, \overline{X}(t)) dW. \end{aligned}$$

Evaluate the integral from 0 to T,

$$\begin{aligned} u(T,\overline{X}(T)) - u(0,X(0)) &= \int_0^T (\bar{a}-a)u_x(t,\overline{X}(t))dt + \int_0^T \frac{\bar{b}^2-b^2}{2}u_{xx}(t,\overline{X}(t))dt \\ &+ \int_0^T \bar{b}(t,\overline{X}(t))u_x dW. \end{aligned}$$

Take the expected value and use (5.10) to obtain

$$\begin{aligned} E[g(\overline{X}(T)) &- g(X(T))] \\ &= \int_0^T E[(\bar{a}-a)u_x] + \frac{1}{2}E[(\bar{b}^2-b^2)u_{xx}]dt + E\left[\int_0^T \bar{b}u_x dW\right] \\ &= \int_0^T E[(\bar{a}-a)u_x] + \frac{1}{2}E[(\bar{b}^2-b^2)u_{xx}]dt. \end{aligned}$$

The following Lemma 5.9 proves the Theorem. □

Lemma 5.9 *There holds for* $t_n \le t < t_{n+1}$

$$\begin{aligned} f_1(t) &\equiv E[(\bar{a}(t,\overline{X}) - a(t,\overline{X}(t)))u_x(t,\overline{X}(t))] = \mathcal{O}(\Delta t_n), \\ f_2(t) &\equiv E[(\bar{b}^2(t,\overline{X}) - b^2(t,\overline{X}(t)))u_{xx}(t,\overline{X}(t))] = \mathcal{O}(\Delta t_n). \end{aligned}$$

Proof. Since $\bar{a}(t,\overline{X}) = a(t_n,\overline{X}(t_n))$,

$$f_1(t_n) = E[(\bar{a}(t_n,\overline{X}) - a(t_n,\overline{X}(t_n)))u_x(t_n,\overline{X}(t_n))] = 0. \tag{5.11}$$

Provided $|f_1'(t)| \le C$, the initial condition (5.11) implies that $f_1(t) = \mathcal{O}(\Delta t_n)$, for $t_n \le t < t_{n+1}$. Therefore, it remains to show that $|f_1'(t)| \le C$. Let $\alpha(t,x) = -(a(t,x) - a(t_n,\overline{X}(t_n)))u_x(t,x)$, so that $f(t) = E[\alpha(t,\overline{X}(t))]$. Then by Itô 's formula

$$\begin{aligned} \frac{df}{dt} = \frac{d}{dt}E\left[\alpha(t,\overline{X}(t))\right] &= E\left[d\alpha(t,\overline{X}(t))\right]/dt \\ &= E\left[\left(\alpha_t + \bar{a}\alpha_x + \frac{\bar{b}^2}{2}\alpha_{xx}\right)dt + \alpha_x\bar{b}dW\right]/dt \\ &= E\left[\alpha_t + \bar{a}\alpha_x + \frac{\bar{b}^2}{2}\alpha_{xx}\right] \\ &= \mathcal{O}(1). \end{aligned}$$

Therefore there exists a constant C such that $|f'(t)| \leq C$, for $t_n < t < t_{n+1}$, and consequently

$$f_1(t) \equiv E[\big(\bar{a}(t,\overline{X}) - a(t,\overline{X}(t))\big)u_x(t,\overline{X}_t)] = \mathcal{O}(\Delta t_n), \quad \text{for } t_n \leq t < t_{n+1}.$$

Similarly, we can also prove

$$f_2(t) \equiv E[\big(\bar{b}^2(t,\overline{X}) - b^2(t,\overline{X}(t))\big)u_{xx}(t,\overline{X}_t)] = \mathcal{O}(\Delta t_n), \quad \text{for } t_n \leq t < t_{n+1}.$$

□

Example 5.10 Consider the stochastic volatility model,

$$\begin{aligned} dS &= \omega S dt + \sigma S dZ \\ d\sigma &= \alpha\sigma dt + v\sigma dW \end{aligned} \tag{5.12}$$

where Z and W are Brownian motions with correlation coefficient ρ, i.e. $E[dZdW] = \rho dt$. We can construct Z and W from the independent W_1 and W_2 by

$$W = W_1, \quad Z = \rho W_1 + \sqrt{1-\rho^2}\, W_2.$$

□

Exercise 5.11 In the risk neutral formulation a stock price solves the stochastic differential equation

$$dS = rSdt + \sigma S dW(t),$$

with constant interest rate r and volatility σ.

1. Show that

$$S(T) = S(0)e^{rT - \frac{\sigma^2}{2}T + \sigma W(T)}. \tag{5.13}$$

2. Use equation (5.13) to simulate the price

$$f(0, S(0)) = e^{-rT}E[\ \max\ (S(T) - K, 0)\ ]$$

of an European call option by a Monte-Carlo method.

3. Compute also the corresponding $\Delta = \partial f(0,S)/\partial S$ by approximating with a difference quotient and determine a good choice of your approximation of $"\partial S"$.

4. Estimate the accuracy of your results. Suggest a better method to solve this problem.

Exercise 5.12 Assume that a system of stocks solves

$$\frac{dS_i}{S_i(t)} = rdt + \sum_{j=1}^{d} \sigma_{ij} dW_j(t) \qquad i = 1, ..., d$$

where W_j are independent Brownian motions.

1. Show that
$$S_i(T) = S(0)e^{rT+\sum_{j=1}^{d}(\sigma_{ij}W_j(T)-\frac{1}{2}\sigma_{ij}^2 T)}.$$

2. Let $S_{av} \equiv \sum_{i=1}^{d} S_i/d$ and simulate the price of the option above with $S(T)$ replaced by $S_{av}(T)$. Estimate the accuracy of your results. Can you find a better method to solve this problem?

Exercise 5.13 [An example of variance reduction] Consider the computation of a call option on an index Z,

$$\pi_t = e^{-r(T-t)} E[\max(Z(T) - K, 0)], \tag{5.14}$$

where Z is the average of d stocks,

$$Z(t) \equiv \frac{1}{d}\sum_{i=1}^{d} S_i(t)$$

and

$$dS_i(t) = rS_i(t)dt + \sigma_i dW_i(t), \quad i = 1, \ldots, d$$

with volatilities

$$\sigma_i \equiv 0.2 * (2 + \sin(i)) \quad i = 1, \ldots, d.$$

The correlation between Wiener processes is given by

$$E[dW_i(t)dW_{i'}(t)] = exp(-2\ |i - i'|/d))dt \quad 1 \le i, i' \le d.$$

The goal of this exercise is to experiment with two different variance reduction techniques, namely the antithetic variates and the control variates.

From now on we take $d = 10$, $r = 0.04$ and $T = 0.5$ in the example above.

(a) Implement a Monte Carlo approximation with for the value in (5.14). Estimate the statistical error. Choose a number of realizations such that the estimate for the statistical error is less than 1% of the value we want to approximate.

(b) Same as (a) but using antithetic variates. The so called *antithetic variates* technique reduces the variance in a sample estimator $\mathcal{A}(M;Y)$ by using another estimator $\mathcal{A}(M;Y')$ with the same expectation as the first one, but which is negatively correlated with the first. Then, the improved estimator is $\mathcal{A}(M;\frac{1}{2}(Y+Y'))$. Here, the choice of Y and Y' relates to the Wiener process W and its reflection along the time axis, $-W$, which is also a Wiener process , i.e.

$$\pi_t \approx \frac{1}{M}\sum_{j=1}^{M} \frac{\{\max(Z(W(T,\omega_j)) - K, 0) + \max(Z(-W(T,\omega_j)) - K, 0)\}}{2}.$$

(c) Same as (a) but using control variates to reduce the variance. The control variates technique is based on the knowledge of an estimator Y'', positively correlated with Y, whose expected value $E[Y'']$ is known and relatively close to the desired $E[Y]$, yielding $Y - Y'' + E[Y'']$ as an improved estimator.

For the application of control variates to (5.14) use the geometric average

$$\hat{Z}(t) \equiv \{\prod_{i=1}^{d} S_i(t)\}^{\frac{1}{d}},$$

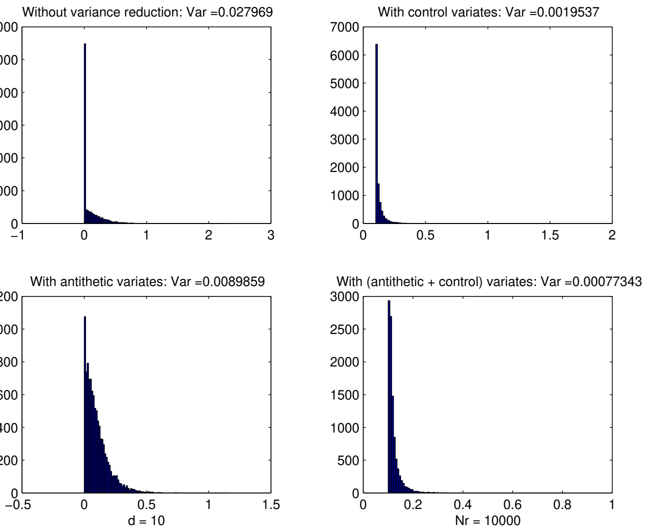
compute

$$\hat{\pi}_t = e^{-r(T-t)} E[\max(\hat{Z}(T) - K, 0)]$$

exactly (hint: find a way to apply Black-Scholes formula). Then approximate

$$\pi_t \approx \hat{\pi}_t + \frac{1}{M}\sum_{j=1}^{M} \left\{\max(Z(W(T,\omega_j)) - K, 0) - \max(\hat{Z}(W(T,\omega_j)) - K, 0)\right\}.$$

(d) Discuss the results from (a-c). Does it pay to use variance reduction?



Without variance reduction: Var =0.027969
With control variates: Var =0.0019537
With antithetic variates: Var =0.0089859
With (antithetic + control) variates: Var =0.00077343
d = 10
Nr = 10000

Chapter 6

Finite Difference Methods

This section introduces finite difference methods for approximation of partial differential equations. We first apply the finite difference method to a partial differential equation for a financial option problem, which is more efficiently computed by partial differential methods than Monte Carlo techniques. Then we discuss the fundamental Lax Equivalence Theorem, which gives the basic understanding of accuracy and stability for approximation of differential equations.

6.1 American Options

Assume that the stock value, S(t), evolves in the risk neutral formulation by the Itô geometric Brownian motion

$$dS = rSdt + \sigma SdW.$$

An American put option is a contract that gives the possibility to sell a stock for a fixed price K up to time T. Therefore the derivation of option values in Chapter 5 shows that European and American options have the formulations:

1. The price of an European put option is

$$f(t,s) \equiv E[\ e^{-r(T-t)} \max(K - S(T), 0)|\ S(t) = s\ ].$$

2. The price of an American option is obtained by maximizing over all sell time τ strategies, which depend on the stock price up to the sell

time,

$$f_A(t,s) \equiv \max_{t\le\tau\le T} E[\ e^{-r(\tau-t)}\max(K-S(\tau),0)|\ S(t)=s\ ].$$

How to find the optimal selling strategy for an American option? Assume that selling is only allowed at the discrete time levels $0, \Delta t, 2\Delta t, \ldots, T$. Consider the small time step $(T-\Delta t, T)$. By assumption the option is not sold in the step. Therefore the European value $f(t,s)$ holds, where $f(T,s) = \max(K-s,0)$ and for $T-\Delta t < t < T$

$$f_t + rSf_S + \frac{1}{2}\sigma^2 S^2 f_{SS} = rf. \tag{6.1}$$

If, for a fixed stock price $s = S(T-\Delta t)$, there holds $f(T-\Delta t, s) < \max(K-s,0)$ then keeping the option gives the expected value $f(T-\Delta t, s)$ which is clearly less than the value $\max(K-s,0)$ obtained by selling at time $T-\Delta t$. Therefore it is optimal to sell if $f(T-\Delta t,s) < \max(K-s,0) \equiv f_F$. Modify the initial data at $t = T-\Delta t$ to $\max(f(T-\Delta t,s), f_F)$ and repeat the step (6.1) for $(T-2\Delta t, T-\Delta t)$ and so on. The price of the American option is obtained as the limit of this solution as $\Delta t \to 0$.

Exercise 6.1 A corresponding Monte Carlo method based on 2 requires simulation of expected values $E[e^{-r\tau}\max(K-S(\tau),0)]$ for many different possible selling time strategies τ until an approximation of the maximum values is found. Try to suggest a strategy for this optimization procedure and compare the efficiency with the partial differential formulation.

Note that the optimal selling strategy

$$\tau = \tau^* = \inf_v\{v : t \le v \le T, f_A(v, S(v)) = \max(K-S(v),0)\}$$

for the American option, which is a function of f_A, seems expensive to evaluate by Monte Carlo technique, but is obtained directly in the partial differential formulation above and below. This technique is a special case of the so called dynamic programming method, which we shall study systematically for general optimization problems in a later Chapter, cf. also the last example in Chapter 1.

Here and in Exercise 6.2 is a numerical method to determine the value of an American option:

(1) Discretize the computational domain $[0,T]\times[s_0,s_1]$ and let

$$f(n\Delta t, i\Delta S)\simeq \bar f_{n,i}, \quad \bar f_{N,i}=\max(K-i\Delta S,0).$$

(2) Use the forward Euler and central difference method for the equation (6.1)

$$\begin{array}{ll} f_t\simeq \frac{\bar f_{n,i}-\hat f_{n-1,i}}{\Delta t} & f_S\simeq \frac{\bar f_{n,i+1}-\bar f_{n,i-1}}{2\Delta S}\\ f_{SS}\simeq \frac{\bar f_{n,i+1}-2\bar f_{n,i}+\bar f_{n,i-1}}{(\Delta S)^2} & f\simeq \bar f_{n,i}. \end{array}$$

(3) Make a Black-Scholes prediction for each time step

$$\begin{aligned} \hat f_{n-1,i} &= \bar f_{n,i}(1-r\Delta t-\sigma^2 i^2\Delta t)+\bar f_{n,i+1}(\frac{1}{2}ri\Delta t+\frac{1}{2}\sigma^2 i^2\Delta t)\\ &+ \bar f_{n,i-1}(-\frac{1}{2}ri\Delta t+\frac{1}{2}\sigma^2 i^2\Delta t). \end{aligned}$$

(4) Compare the prediction with selling by letting

$$\bar f_{n-1,i}=\max\left(\hat f_{n-1,i},\max(K-i\Delta S,0)\right),$$

and go to the next time Step 3 by decreasing n by 1.

Exercise 6.2 The method above needs in addition boundary conditions at $S=s_0$ and $S=s_1$ for $t<T$. How can s_0, s_1 and these conditions be choosen to yield a good approximation?

Exercise 6.3 Give a trinomial tree interpretation of the finite difference scheme

$$\begin{aligned} \bar f_{n+1,i} &= \bar f_{n,i}(1+r\Delta t+\sigma^2 i^2\Delta t)+\bar f_{n,i+1}(-\frac{1}{2}ri\Delta t-\frac{1}{2}\sigma^2 i^2\Delta t)\\ &+ \bar f_{n,i-1}(\frac{1}{2}ri\Delta t-\frac{1}{2}\sigma^2 i^2\Delta t), \end{aligned}$$

for Black-Scholes equation of an European option. Binomial and trinomial tree approximations are frequent in the finance economy literature, cf. [J. Hull].

Let us now study general finite difference methods for partial differential equations. The motivation to introduce general finite difference methods in contrast to study only the binomial and trinomial tree methods is that higher order methods, such as the Crank-Nicolson method below, are more efficient to solve e.g. (6.1).

The error for the binomial and the trinomial tree method applied to the partial differential equation (6.1) for a European option is $\varepsilon = \mathcal{O}(\Delta t + (\Delta s)^2)$, which is clearly the same for the related forward and backward Euler methods. The work is then $\mathcal{A} = \mathcal{O}((\Delta t \Delta s)^{-1})$, so that $\mathcal{A} = \mathcal{O}(\varepsilon^{-3/2})$. For the Crank-Nicolsen method the accuracy is $\varepsilon = \mathcal{O}((\Delta t)^2 + (\Delta s)^2)$ and the work is still $\mathcal{A} = \mathcal{O}((\Delta t \Delta s)^{-1})$, which implies the improved bound $\mathcal{A} = \mathcal{O}(\varepsilon^{-1})$. For a general implicit method with a smooth exact solution in $[0, T] \times \mathbb{R}^d$ the accuracy is $\varepsilon = \mathcal{O}((\Delta t)^q + (\Delta s)^p)$ with the miminal work (using e.g. the multigrid method) $\mathcal{A} = \mathcal{O}(\frac{q^2}{\Delta t}(\frac{p^2}{\Delta s})^d)$, which gives $\mathcal{A} = \mathcal{O}(\frac{q^2}{\varepsilon^{1/q}}(\frac{p^2}{\varepsilon^{1/p}})^d)$. In the next section we derive these error estimates for some model problems.

6.2 Lax Equivalence Theorem

Lax equivalence theorem defines the basic concepts for approximation of linear well posed differential equations. Here, well posed means that the equation is solvable for data in a suitable function space and that the solution operator is bounded. We will first formally state the result without being mathematically precise with function spaces and norms. Then we present two examples with proofs based on norms and functions spaces.

The ingredients of Lax Equivalence Theorem 6.4 are:

(0) an exact solution u, satisfying the *linear well posed equation* $Lu = f$, and an approximation u_h, obtained from $L_h u_h = f_h$;

(1) *stability*, the approximate solution operators $\|L_h^{-1}\|$ are uniformly bounded in h and the exact solution operator $\|L^{-1}\|$ is bounded;

(2) *consistency*, $f_h \to f$ and $L_h u \to Lu$ as the mesh size $h \to 0$; and

(3) *convergence*, $u_h \to u$ as the mesh size $h \to 0$.

Theorem 6.4 *The combination of stability and consistency is equivalent to convergence.*

The idea of the proof. To verify convergence, consider the identity

$$u - u_h = L_h^{-1}[\ L_h u - L_h u_h\ ] \stackrel{Step(0)}{=} L_h^{-1}[\ (L_h u - Lu) + (f - f_h)\ ].$$

Stability implies that L_h^{-1} is bounded and consistency implies that

$$L_h u - Lu \to 0 \text{ and } f - f_h \to 0,$$

and consequently the convergence holds

$$\begin{aligned} \lim_{h\to 0}(u - u_h) &= \lim_{h\to 0} L_h^{-1}[\ (L_h u - Lu) + (f - f_h)\ ] \\ &= 0. \end{aligned}$$

Clearly, consistency is necessary for convergence. Example 6.7, below, indicates that also stability is necessary. □

Let us now more precisely consider the requirements and norms to verify stability and consistency for two concrete examples of ordinary and partial differential equations.

Example 6.5 Consider the forward Euler method for the ordinary differential equation

$$\begin{aligned} &u'(t) = Au(t) \quad 0 < t < 1, \\ &u(0) = u_0. \end{aligned} \tag{6.2}$$

Verify the conditions of stability and consistency in Lax Equivalence Theorem.

Solution. For a given partition, $0 = t_0 < t_1 < ... < t_N = 1$, with $\Delta t = t_{n+1} - t_n$, let

$$\begin{aligned} u_{n+1} &\equiv (I + \Delta t A)u_n \\ &= G^n u_0 \qquad \text{where } G = (I + \Delta t A). \end{aligned}$$

Then:

(1) Stability means $|G^n| + |H^n| \le e^{Kn\Delta t}$ for some K, where $|\cdot|$ denotes the matrix norm $|F| \equiv \sup_{\{v\in\mathbb{R}^n : |v|\le 1\}} |Fv|$ with the Euclidean norm $|w| \equiv \sqrt{\sum_i w_i^2}$ in $\mathbb{R}^n$.

(2) Consistency means $|(G-H)v| \leq C(\Delta t)^{p+1}$, where $H = e^{\Delta t A}$ and p is the order of accuracy. In other words, the consistency error $(G-H)v$ is the local approximation error after one time step with the same initial data v.

This stability and consistency imply the convergence

$$\begin{aligned}
\mid u_n - u(n\Delta t) \mid &= \mid (G^n - H^n)u_0 \mid \\
&= \mid (G^{n-1} + G^{n-2}H + ... + GH^{n-2} + H^{n-1})(G-H)u_0 \mid \\
&\leq \mid G^{n-1} + G^{n-2}H + ... + GH^{n-2} + H^{n-1} || (G-H)u_0 \mid \\
&\leq C(\Delta t)^{p+1} n | \; u_0 \; | e^{Kn\Delta t} \\
&\leq C'(\Delta t)^p,
\end{aligned}$$

with the convergence rate $\mathcal{O}(\Delta t^p)$. For example, $p = 1$ in case of the Euler method and $p = 2$ in case of the trapezoidal method. □

Example 6.6 Consider the heat equation

$$\begin{aligned}
u_t &= u_{xx} \quad t > 0, \\
u(0) &= u_0.
\end{aligned} \tag{6.3}$$

Verify the stability and consistency conditions in Lax Equivalence Theorem.

Solution. Apply the Fourier transform to equation (6.3),

$$\hat{u}_t = -\omega^2 \hat{u}$$

so that

$$\hat{u}(t, \omega) = e^{-t\omega^2} \hat{u}_0(\omega).$$

Therefore $\hat{H} = e^{-\Delta t \omega^2}$ is the exact solution operator for one time step, i.e. $\hat{u}(t + \Delta t) = \hat{H}\hat{u}(t)$. Consider the difference approximation of (6.3)

$$\frac{u_{n+1,i} - u_{n,i}}{\Delta t} = \frac{u_{n,i+1} - 2u_{n,i} + u_{n,i-1}}{\Delta x^2},$$

which shows

$$u_{n+1,i} = u_{n,i}\left(1 - \frac{2\Delta t}{\Delta x^2}\right) + \frac{\Delta t}{\Delta x^2}\left(u_{n,i+1} + u_{n,i-1}\right),$$

where $u_{n,i} \simeq u(n\Delta t, i\Delta x)$. Apply the Fourier transform to obtain

$$\begin{aligned}
\hat{u}_{n+1} &= \left[\left(1-\frac{2\Delta t}{\Delta x^2}\right)+\frac{\Delta t}{\Delta x^2}\left(e^{j\Delta x\omega}+e^{-j\Delta x\omega}\right)\right]\hat{u}_n \\
&= \left[1-2\frac{\Delta t}{\Delta x^2}+2\frac{\Delta t}{\Delta x^2}\cos(\Delta x\omega)\right]\hat{u}_n \\
&= \hat{G}\hat{u}_n \qquad \left(\text{ Let } \hat{G} \equiv 1-2\frac{\Delta t}{\Delta x^2}+2\frac{\Delta t}{\Delta x^2}\cos(\Delta x\omega)\right) \\
&= \hat{G}^{n+1}\hat{u}_0.
\end{aligned}$$

1. We have

$$\begin{aligned}
2\pi\|u_n\|^2_{L^2} &= \|\hat{u}_n\|^2_{L^2} \qquad \text{(by Parseval's formula)} \\
&= \|\hat{G}^n\hat{u}_0\|^2_{L^2} \\
&\le \sup_\omega |\hat{G}^n|^2 \ \|\hat{u}_0\|^2_{L^2}.
\end{aligned}$$

Therefore the condition

$$\|\hat{G}^n\|_{L^\infty} \le e^{Kn\Delta t} \tag{6.4}$$

implies L^2-stability.

2. We have

$$2\pi\|u_1-u(\Delta t)\|^2_{L^2} = \|\hat{G}\hat{u}_0-\hat{H}\hat{u}_0\|^2_{L^2},$$

where u_1 is the approximate solution after one time step. Let $\lambda \equiv \frac{\Delta t}{\Delta x^2}$, then we obtain

$$\begin{aligned}
|(\hat{G}-\hat{H})\hat{u}_0| &= |\left(1-2\lambda+2\lambda\cos\Delta x\omega - e^{-\Delta t\omega^2}\right)\hat{u}_0| \\
&= \mathcal{O}(\Delta t^2)\omega^4|\hat{u}_0|,
\end{aligned}$$

since for $0 \le \Delta t\omega^2 \equiv x \le 1$

$$\begin{aligned}
|1-2\lambda \ &+ \ 2\lambda\cos\sqrt{x/\lambda}-e^{-x}| \\
&= \left(1-2\lambda+2\lambda\left(1-\frac{x}{2\lambda}+\mathcal{O}(x^2)\right)-\left(1-x+\mathcal{O}(x^2)\right)\right) \\
&\le Cx^2 = C(\Delta t)^2\omega^4,
\end{aligned}$$

and for $1 < \Delta t\omega^2 = x$

$$|1-2\lambda+2\lambda\cos\sqrt{x/\lambda}-e^{-x}| \le C = C\frac{(\Delta t)^2\omega^4}{x^2} \le C(\Delta t)^2\omega^4.$$

Therefore the consistency condition reduces to

$$\begin{aligned}\| \, (\hat{G} - \hat{H})\hat{u}_0 \, \| &\leq \|K\Delta t^2 \omega^4 \hat{u}_0\| \\ &\leq K\Delta t^2 \|\partial_{xxxx} u_0\|_{L^2}.\end{aligned} \tag{6.5}$$

3. The stability (6.4) holds if

$$\|\hat{G}\|_{L^\infty} \equiv \sup_\omega |\hat{G}(\omega)| = \max_\omega |1 - 2\lambda + 2\lambda \cos \Delta x\omega| \leq 1, \tag{6.6}$$

which requires

$$\lambda = \frac{\Delta t}{\Delta x^2} \leq \frac{1}{2}. \tag{6.7}$$

The L^2-stability condition (6.6) is called the von Neuman stability condition.

4. Convergence follows by the estimates (6.5), (6.6) and $\|\hat{H}\|_{L^\infty} \leq 1$

$$\begin{aligned} 2\pi \| \, u_n - u(n\Delta t) \, \|^2_{L^2} &= \| \, (\hat{G}^n - \hat{H}^n)\hat{u}_0 \, \|^2_{L^2} \\ &= \| \, (\hat{G}^{n-1} + \hat{G}^{n-2}\hat{H} + ... + \hat{H}^{n-1})(\hat{G} - \hat{H})\hat{u}_0 \, \|^2_{L^2} \\ &\leq \| \, \hat{G}^{n-1} + \hat{G}^{n-2}\hat{H} + ... + \hat{H}^{n-1}\|^2_{L^\infty} \|(\hat{G} - \hat{H})\hat{u}_0 \, \|^2_{L^2} \\ &\leq (Kn(\Delta t)^2)^2 \leq (KT\Delta t)^2, \end{aligned}$$

and consequently the convergence rate is $\mathcal{O}(\Delta t)$. □

Let us study the relations between the operators G and H for the simple model problem

$$\begin{aligned} u' + \lambda u &= 0 \\ u(0) &= 1 \end{aligned}$$

with an approximate solution $u_{n+1} = r(x)u_n$ (where $x = \lambda\Delta t$):

(1) the exact solution satisfies

$$r(x) = e^{-\lambda\Delta t} = e^{-x},$$

(2) the forward Euler method

$$\frac{u_{n+1} - u_n}{\Delta t} + \lambda u_n = 0 \quad \Rightarrow \quad r(x) = 1 - x,$$

(3) the backward Euler method

$$\frac{u_{n+1} - u_n}{\Delta t} + \lambda u_{n+1} = 0 \quad \Rightarrow \quad r(x) = (1 + x)^{-1},$$

(4) the trapezoidal method

$$\frac{u_{n+1} - u_n}{\Delta t} + \frac{\lambda}{2}(u_n + u_{n+1}) = 0 \quad \Rightarrow \quad r(x) = \left(1 + \frac{x}{2}\right)^{-1} \left(1 - \frac{x}{2}\right),$$

and

(5) the Lax-Wendroff method

$$u_{n+1} = u_n - \Delta t \lambda u_n + \frac{1}{2}\Delta t^2 \lambda^2 u_n \quad \Rightarrow \quad r(x) = 1 - x + \frac{1}{2}x^2.$$

The consistence $|e^{-\lambda \Delta t} - r(\lambda \Delta t)| = \mathcal{O}(\Delta t^{p+1})$ holds with $p = 1$ in case 2 and 3, and $p = 2$ in case 4 and 5. The following stability relations hold:

(1) $r(x) \leq 1$ for $x \geq 0$ in case 1, 3 and 4.

(2) $r(x) \to 0$ as $x \to \infty$ in case 1 and 3.

(3) $r(x) \to 1$ as $x \to \infty$ in case 4.

Property (1) shows that for $\lambda > 0$ case 3 and 4 are unconditionally stable. However Property (2) and (3) refine this statement and imply that only case 3 has the same damping behavior for large λ as the exact solution. Although the damping Property (2) is not necessary to prove convergence it is advantegous to have for proplems with many time scales, e.g. for a system of equations (6.2) where A has eigenvalues $\lambda_i \leq 1, \ i = 1, \ldots, N$ and some $\lambda_j \ll -1$, (why?).

The unconditionally stable methods, e.g. case 3 and 4, are in general more efficient to solve parabolic problems, such as the Black-Scholes equation (6.1), since they require for the same accuracy fewer time steps than the explicit methods, e.g. case 2 and 5. Although the work in each time step for the unconditionally stable methods may be larger than for the explicit methods.

Exercise 6.7 Show by an example that $\|u_n\|_{L^2}^2 \to \infty$ if for some ω there holds $|\hat{G}(\omega)| > 1$, in Example 6.6, i.e. the von Neumann stability condition does not hold.

Chapter 7

The Finite Element Method and Lax-Milgram's Theorem

This section presents the finite element method, including adaptive approximation and error estimates, together with the basic theory for elliptic partial differential equations. The motivation to introduce finite element methods is the computational simplicity and efficiency for construction of stable higher order discretizations for elliptic and parabolic differential equations, such as the Black and Scholes equation, including general boundary conditions and domains. Finite element methods require somewhat more work per degree of freedom as compared to finite difference methods on a uniform mesh. On the other hand, construction of higher order finite difference approximations including general boundary conditions or general domains is troublesome.

In one space dimension such an elliptic problem can, for given functions $a, f, r : (0,1) \to \mathbf{R}$, take the form of the following equation for $u : [0,1] \to \mathbf{R}$,

$$\begin{aligned} &(-au')' + ru = f && \text{on } \ (0,1) \\ &u(x) = 0 && \text{for } x = 0, \ x = 1, \end{aligned} \tag{7.1}$$

where $a > 0$ and $r \geq 0$. The basic existence and uniqueness result for general elliptic differential equations is based on Lax-Milgram's Theorem, which we will describe in section 7.3. We shall see that its stability properties, based on so called energy estimates, is automatically satisfied for finite element methods in contrast to finite difference methods.

Our goal, for a given tolerence TOL, is to find an approximation u_h of (7.1) satisfying

$$\|u - u_h\| \leq \text{TOL},$$

using few degrees of freedoom by adaptive finite element approximation. Adaptive methods are based on:

(1) an automatic mesh generator,

(2) a numerical method (e.g. the finite element method),

(3) a refinement criteria (e.g. a posteriori error estimation), and

(4) a solution algorithm (e.g. the multigrid method).

7.1 The Finite Element Method

A derivation of the finite element method can be divided into:

(1) variational formulation in an infinite dimensional space V,

(2) variational formulation in a finite dimensional subspace, $V_h \subset V$,

(3) choice of a basis for V_h, and

(4) solution of the discrete system of equations.

Step 1. *Variational formulation in an infinite dimensional space, V.* Consider the following Hilbert space,

$$V = \left\{ v : (0,1) \to \mathbf{R} \ : \ \int_0^1 \left(v^2(x) + (v'(x))^2 \right) dx < \infty, \ v(0) = v(1) = 0 \right\}.$$

Multiply equation (7.1) by $v \in V$ and integrate by parts to get

$$\begin{aligned} \int_0^1 fv \; dx &= \int_0^1 ((-au')' + ru)v \; dx \\ &= [-au'v]_0^1 + \int_0^1 (au'v' + ruv) \; dx \qquad (7.2) \\ &= \int_0^1 (au'v' + ruv) \; dx. \end{aligned}$$

Therefore the variational formulation of (7.1) is to find $u \in V$ such that

$$A(u,v) = L(v) \qquad \forall v \in V, \tag{7.3}$$

where

$$\begin{aligned} A(u,v) &= \int_0^1 (au'v' + ruv)\; dx, \\ L(v) &= \int_0^1 fv\; dx. \end{aligned}$$

Remark 7.1 The integration by parts in (7.2) shows that a smooth solution of equation (7.1) satisfies the variational formulation (7.3). For a solution of the variational formulation (7.3) to also be a solution of the equation (7.1), we need additional conditions on the regularity of the functions a, r and f so that u'' is continuous. Then the following integration by parts yields, as in (7.2),

$$0 = \int_0^1 (au'v' + ruv - fv)\; dx = \int_0^1 (-(au')' + ru - f)v\; dx.$$

Since this holds for all $v \in V$, it implies that

$$-(au')' + ru - f = 0,$$

provided $-(au')' + ru - f$ is continuous. $\square$

Step 2. *Variational formulation in the finite dimensional subspace, V_h.* First divide the interval $(0,1)$ into $0 = x_0 < x_2 < ... < x_{N+1} = 1$, i.e. generate the mesh. Then define the space of continuous piecewise linear functions on the mesh with zero boundary conditions

$$\begin{aligned} V_h = \{v \in V \;\; : \;\; & v(x)\mid_{(x_i,x_{i+1})} = c_i x + d_i, \text{ i.e. } v \text{ is linear on } (x_i, x_{i+1}), i = 0, \cdots, N \\ & \text{and } v \text{ is continuous on } (0,1)\}. \end{aligned}$$

The variational formulation in the finite dimensional subspace is to find $u_h \in V_h$ such that

$$A(u_h, v) = L(v) \qquad \forall v \in V_h. \tag{7.4}$$

The function u_h is a finite element solution of the equation (7.1). Other finite element solutions are obtained from alternative finite dimensional subspaces, e.g. based on piecewise quadratic approximation.

Step 3. *Choose a basis for V_h.*

Let us introduce the basis functions $\phi_i \in V_h$, for $i = 1, ..., N$, defined by

$$\phi_i(x_j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases} \tag{7.5}$$

A function $v \in V_h$ has the representation

$$v(x) = \sum_{i=1}^{N} v_i \phi_i(x),$$

where $v_i = v(x_i)$, i.e. each $v \in V_h$ can be written in a unique way as a linear combination of the basis functions ϕ_i.

Step 4. *Solve the discrete problem (7.4).*

Using the basis functions ϕ_i, for $i = 1, ..., N$ from Step 3, we have

$$u_h(x) = \sum_{i=1}^{N} \xi_i \phi_i(x),$$

where $\xi = (\xi_1, ..., \xi_N)^T \in \mathbf{R}^N$, and choosing $v = \phi_j$ in (7.4), we obtain

$$\begin{aligned} L(\phi_j) &= A(u_h, \phi_j) \\ &= A(\sum_i \phi_i \xi_i, \phi_j) = \sum_i \xi_i A(\phi_i, \phi_j), \end{aligned}$$

so that $\xi \in \mathbf{R}^N$ solves the linear system

$$\tilde{A}\xi = \tilde{L}, \tag{7.6}$$

where

$$\begin{aligned} \tilde{A}_{ji} &= A(\phi_i, \phi_j), \\ \tilde{L}_j &= L(\phi_j). \end{aligned}$$

The $N \times N$ matrix $\tilde{A}$ is called the stiffness matrix and the vector $\tilde{L} \in \mathbf{R}^N$ is called the load vector.

Example 5.1 Consider the following two dimensional problem,

$$\begin{aligned} -div(k\nabla u) + ru &= f \quad \text{in } \Omega \subset \mathbb{R}^2 \\ u &= g_1 \quad \text{on } \Gamma_1 \\ \frac{\partial u}{\partial n} &= g_2 \quad \text{on } \Gamma_2, \end{aligned} \tag{7.7}$$

where $\partial\Omega = \Gamma = \Gamma_1 \cup \Gamma_2$ and $\Gamma_1 \cap \Gamma_2 = \emptyset$. The variational formulation has the following form.

1. Variational formulation in the infinite dimensional space.

Let

$$V_g = \left\{ v(x) : \int_\Omega (v^2(x) + |\nabla v(x)|^2)\ dx < \infty, v|_{\Gamma_1} = g \right\}.$$

Take a function $v \in V_0$, i.e. $v = 0$ on Γ_1, then by (7.7)

$$\begin{aligned} \int_\Omega fv\ dx &= -\int_\Omega div(k\nabla u)v\ dx + \int_\Omega ruv\ dx \\ &= \int_\Omega k\nabla u \cdot \nabla v\ dx - \int_{\Gamma_1} k\frac{\partial u}{\partial n}v\ ds - \int_{\Gamma_2} k\frac{\partial u}{\partial n}v\ ds + \int_\Omega ruv\ dx \\ &= \int_\Omega k\nabla u \cdot \nabla v\ dx - \int_{\Gamma_2} kg_2 v\ ds + \int_\Omega ruv\ dx. \end{aligned}$$

The variational formulation for the model problem (7.7) is to find $u \in V_{g_1}$ such that

$$A(u, v) = L(v) \qquad \forall v \in V_0, \tag{7.8}$$

where

$$\begin{aligned} A(u, v) &= \int_\Omega (k\nabla u \cdot \nabla v + ruv)\ dx, \\ L(v) &= \int_\Omega fv\ dx + \int_{\Gamma_2} kg_2 v ds. \end{aligned}$$

2. Variational formulation in the finite dimensional space.

Assume for simplicity that Ω is a polygonial domain which can be divided into a triangular mesh $T_h = \{K_1, ...K_N\}$ of non overlapping triangles K_i and let $h = \max_i$(length of longest side of K_i). Assume also that the boundary function g_1 is continuous and that its restriction to each edge $K_i \cap \Gamma_1$ is a linear function. Define

$$\begin{aligned} V_0^h &= \{v \in V_0 \ : \ v|_{K_i} \text{ is linear } \forall K_i \in T_h, \ v \text{ is continuous on } \Omega\}, \\ V_{g_1}^h &= \{v \in V_{g_1} \ : \ v|_{K_i} \text{ is linear } \forall K_i \in T_h, \ v \text{ is continuous on } \Omega\}, \end{aligned}$$

and the finite element method is to find $u_h \in V_{g_1}^h$ such that

$$A(u_h, v) = L(v), \qquad \forall v \in V_0^h. \tag{7.9}$$

3. Choose a basis for V_0^h.
As in the one dimensional problem, choose the basis $\phi_j \in V_0^h$ such that

$$\phi_j(x_i) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \qquad j = 1, 2, ..., N,$$

where $x_i, \ i = 1, \ldots, N$, are the vertices of the triangulation.

4. Solve the discrete system.
Let

$$u_h(x) = \sum_{i=1}^{N} \xi_i \phi_i(x), \quad \text{and} \quad \xi_i = u_h(x_i).$$

Then (7.9) can be written in matrix form,

$$\tilde{A}\xi = \tilde{L}, \quad \text{where} \quad \tilde{A}_{ji} = A(\phi_i, \phi_j) \ \text{ and } \ \tilde{L}_j = L(\phi_j).$$

□

7.2 Error Estimates and Adaptivity

We shall now study a priori and a posteriori error estimates for finite element methods, where

$$\begin{aligned} \|u - u_h\| &\leq E_1(h, u, f) \quad \text{is an a priori error estimate,} \\ \|u - u_h\| &\leq E_2(h, u_h, f) \quad \text{is an a posteriori error estimate.} \end{aligned}$$

Before we start, let us study the following theorem, which we will prove later,

Theorem 7.2 (Lax-Milgram) *Let V be a Hilbert space with norm $\|\cdot\|_V$ and scalar product $(\cdot,\cdot)_V$ and assume that A is a bilinear functional and L is a linear functional that satisfy:*

(1) A is symmetric, i.e. $A(v,w) = A(w,v) \quad \forall v, w \in V$;

(2) A is V-elliptic, i.e. $\exists\ \alpha > 0$ such that $A(v,v) \geq \alpha \|v\|_V^2 \quad \forall v \in V$;

(3) A is continuous, i.e. $\exists\ C \in \mathbb{R}$ such that $|A(v,w)| \leq C\|v\|_V \|w\|_V$; and

(4) L is continuous, i.e. $\exists\ \Lambda \in \mathbb{R}$ such that $|L(v)| \leq \Lambda \|v\|_V \quad \forall v \in V$.

Then there is a unique function $u \in V$ such that $A(u,v) = L(v) \quad \forall v \in V$, and the stability estimate $\|u\|_V \leq \Lambda/\alpha$ holds.

7.2.1 An A Priori Error Estimate

The approximation property of the space V_h can be characterized by

Lemma 7.3 *Suppose V_h is the piecewise linear finite element space (7.4), which discretizes the functions in V, defined on $(0,1)$, with the interpolant $\pi : V \to V_h$ defined by*

$$\pi v(x) = \sum_{i=1}^{N} v(x_i)\phi_i(x), \tag{7.10}$$

where $\{\phi_i\}$ is the basis (7.5) of V_h. Then

$$\begin{aligned} \|(v - \pi v)'\|_{L^2(0,1)} &\leq \sqrt{\int_0^1 h^2 v''(x)^2 \ dx} \leq Ch, \\ \|v - \pi v\|_{L^2(0,1)} &\leq \sqrt{\int_0^1 h^4 v''(x)^2 \ dx} \leq Ch^2, \end{aligned} \tag{7.11}$$

where $h = \max_i \ (x_{i+1} - x_i)$.

Proof. Take $v \in V$ and consider first (7.11) on an interval (x_i, x_{i+1}). By the mean value theorem, there is for each $x \in (x_i, x_{i+1})$ a $\xi \in (x_i, x_{i+1})$ such that $v'(\xi) = (\pi v)'(x)$. Therefore

$$v'(x) - (\pi v)'(x) = v'(x) - v'(\xi) = \int_\xi^x v''(s)ds,$$

so that

$$\begin{aligned}
\int_{x_i}^{x_{i+1}} |v'(x) - (\pi v)'(x)|^2 dx &= \int_{x_i}^{x_{i+1}} (\int_\xi^x v''(s)ds)^2 dx \\
&\leq \int_{x_i}^{x_{i+1}} |x - \xi| \int_\xi^x (v''(s))^2 ds dx \\
&\leq h^2 \int_{x_i}^{x_{i+1}} (v''(s))^2 ds, \qquad (7.12)
\end{aligned}$$

which after summation of the intervals proves (7.11).

Next, we have

$$v(x) - \pi v(x) = \int_{x_i}^x (v - \pi v)'(s)ds,$$

so by (7.12)

$$\begin{aligned}
\int_{x_i}^{x_{i+1}} |v(x) - \pi v(x)|^2 dx &= \int_{x_i}^{x_{i+1}} (\int_{x_i}^x (v - \pi v)'(s)ds)^2 dx \\
&\leq \int_{x_i}^{x_{i+1}} |x - x_i| \int_{x_i}^x ((v - \pi v)')^2(s) ds dx \\
&\leq h^4 \int_{x_i}^{x_{i+1}} (v''(s))^2 ds,
\end{aligned}$$

which after summation of the intervals proves the lemma. □

Our derivation of the a priori error estimate

$$\|u - u_h\|_V \leq Ch,$$

where u and u_h satisfy (7.3) and (7.4), respectivly, uses Lemma 7.3 and a combination of the following four steps:

(1) error representation based on the *ellipticity*

$$\alpha \int_\Omega (v^2(x) + (v'(x))^2)\ dx \leq A(v,v) = \int_\Omega (a(v')^2 + rv^2)\ dx,$$

where $\alpha = inf_{x\in(0,1)}(a(x), r(x)) > 0$,

(2) the *orthogonality*

$$A(u - u_h, v) = 0 \quad \forall v \in V_h,$$

obtained by $V_h \subset V$ and subtraction of the two equations

$$\begin{aligned} A(u,v) &= L(v) \quad \forall v \in V \qquad \text{by (7.3)}, \\ A(u_h,v) &= L(v) \quad \forall v \in V_h \qquad \text{by (7.4)}, \end{aligned}$$

(3) the *continuity*

$$|A(v,w)| \leq C\|v\|_V \|w\|_V \quad \forall v,w \in V,$$

where $C \leq \sup_{x\in(0,1)}(a(x), r(x))$, and

(4) the *interpolation estimates*

$$\begin{aligned} \|(v - \pi v)'\|_{L^2} &\leq Ch, \\ \|v - \pi v\|_{L^2} &\leq Ch^2, \end{aligned} \tag{7.13}$$

where $h = \max\ (x_{i+1} - x_i)$.

To start the proof of an a priori estimate let $e \equiv u - u_h$. Then by Cauchy's inequality

$$\begin{aligned} A(e,e) &= A(e, u - \pi u + \pi u - u_h) \\ &= A(e, u - \pi u) + A(e, \pi u - u_h) \\ &\overset{\text{Step2}}{=} A(e, u - \pi u) \\ &\leq \sqrt{A(e,e)}\sqrt{A(u - \pi u, u - \pi u)}, \end{aligned}$$

so that by division of $\sqrt{A(e,e)}$,

$$\begin{aligned} \sqrt{A(e,e)} &\leq \sqrt{A(u - \pi u, u - \pi u)} \\ &\overset{\text{Step3}}{=} C\|u - \pi u\|_V \\ &\equiv C\sqrt{\|u - \pi u\|^2_{L^2} + \|(u - \pi u)'\|^2_{L^2}} \\ &\overset{\text{Step4}}{\leq} Ch. \end{aligned}$$

Therefore, by Step 1

$$\alpha\|e\|_V^2 \le A(e,e) \le Ch^2,$$

which implies the a priori estimate

$$\|e\|_V \le Ch,$$

where $C = K(u)$. □

7.2.2 An A Posteriori Error Estimate

Example 7.4 Consider the model problem (7.1), namely,

$$\begin{cases} -(au')' + ru = f & \text{in } (0,1), \\ u(0) = u(1) = 0. \end{cases}$$

Then

$$\begin{aligned} \sqrt{A(u-u_h, u-u_h)} &\le C\ \|a^{-\frac{1}{2}}(f - ru_h + a'u_h')h\|_{L^2} \\ &\equiv E(h, u_h, f). \end{aligned} \tag{7.14}$$

Proof. Let $e = u - u_h$ and let $\pi e \in V_h$ be the nodal interpolant of e. We have

$$\begin{aligned} A(e,e) &= A(e, e - \pi e) \qquad \text{(by orthogonality)} \\ &= A(u, e-\pi e) - A(u_h, e - \pi e). \end{aligned}$$

Using the notation $(f,v) \equiv \int_0^1 fv\ dx$, we obtain by integration by parts

$$\begin{aligned} A(e,e) &= (f, e-\pi e) - \sum_{i=1}^{N} \int_{x_i}^{x_{i+1}} (au_h'(e-\pi e)' + ru_h(e-\pi e))\ dx \\ &= (f - ru_h, e - \pi e) - \sum_{i=1}^{N} \left\{ [au_h'(e-\pi e)]_{x_i}^{x_{i+1}} - \int_{x_i}^{x_{i+1}} (au_h')'(e-\pi e)\ dx \right\} \\ &= (f - ru_h + a'u_h', e - \pi e) \quad (\text{ since } \ u_h''|_{(x_i,x_{i+1})} = 0,\ (e-\pi e)(x_i) = 0) \\ &\le \|a^{-\frac{1}{2}}h(f - ru_h + a'u_h')\|_{L^2} \|a^{\frac{1}{2}}h^{-1}(e-\pi e)\|_{L^2}. \end{aligned}$$

Lemma 7.5 implies

$$\sqrt{A(e,e)} \le C\|a^{-\frac{1}{2}}h(f - ru_h + a'u_h')\|_{L^2},$$

which also shows that

$$\|e\|_V \le Ch,$$

where $C = K'(u_h)$. $\square$

Lemma 7.5 *There is a constant C, independent of u and u_h, such that,*

$$\|a^{\frac{1}{2}}h^{-1}(e - \pi e)\|_{L^2} \le C\sqrt{\int_0^1 ae'e'\ dx} \le C\sqrt{A(e,e)}$$

$\square$

Exercise 7.6 Use the interpolation estimates in Lemma 7.3 to prove Lemma 7.5.

7.2.3 An Adaptive Algorithm

We formulate an adaptive algorithm based on the a posteriori error estimate (7.14) as follows:

(1) Choose an initial coarse mesh T_{h_0} with mesh size h_0.

(2) Compute the corresponding FEM solution u_{h_i} in V_{h_i}.

(3) Given a computed solution u_{h_i} in V_{h_i}, with the mesh size h_i,

$$\begin{array}{ll} \text{stop} & \text{if } \ E(h_i, u_{h_i}, f) \le TOL \\ \text{go to step 4} & \text{if } \ E(h_i, u_{h_i}, f) > TOL. \end{array}$$

(4) Determine a new mesh $T_{h_{i+1}}$ with mesh size h_{i+1} such that

$$E(h_{i+1}, u_{h_i}, f) \cong TOL,$$

by letting the error contribution for all elements be approximately constant, i.e.

$$\|a^{-\frac{1}{2}}h(f - ru_h - a'u_h')\|_{L^2(x_i,x_{i+1})} \cong C, \quad i = 1, \ldots, N,$$

then go to Step 2.

7.3 Lax-Milgram's Theorem

Theorem 7.7 *Suppose A is symmetric, i.e. $A(u,v) = A(v,u) \quad \forall u, v \in V$, then (Variational problem) $\Longleftrightarrow$ (Minimization problem) with*

$$\begin{array}{ll} \text{(Var)} & \text{Find } u \in V \text{ such that } A(u,v) = L(v) \quad \forall v \in V, \\ \text{(Min)} & \text{Find } u \in V \text{ such that } F(u) \leq F(v) \quad \forall v \in V, \end{array}$$

where

$$F(w) \;\equiv\; \frac{1}{2}A(w,w) - L(w) \qquad \forall w \in V.$$

Proof. Take $\epsilon \in \mathbb{R}$. Then

$$\begin{aligned} (\Rightarrow) \quad F(u + \epsilon w) &= \frac{1}{2}A(u+\epsilon w, u+\epsilon w) - L(u+\epsilon w) \\ &= \left(\frac{1}{2}A(u,u) - L(u)\right) + \epsilon A(u,w) - \epsilon L(w) + \frac{1}{2}\epsilon^2 A(w,w) \\ &\geq \left(\frac{1}{2}A(u,u) - L(u)\right) \quad \left(\text{since } \frac{1}{2}\epsilon^2 A(w,w) \geq 0 \text{ and } A(u,w) = L(w)\right) \\ &= F(u). \end{aligned}$$

($\Leftarrow$) Let $g(\epsilon) = F(u + \epsilon w)$, where $g : \mathbf{R} \to \mathbf{R}$. Then

$$0 = g'(0) = 0 \cdot A(w,w) + A(u,w) - L(w) = A(u,w) - L(w).$$

Therefore

$$A(u,w) = L(w) \quad \forall w \in V.$$

$\square$

Theorem 7.8 (Lax-Milgram) *Let V be a Hilbert space with norm $\|\cdot\|_V$ and scalar product $(\cdot,\cdot)_V$ and assume that A is a bilinear functional and L is a linear functional that satisfy:*

(1) A is symmetric, i.e. $A(v,w) = A(w,v) \quad \forall v, w \in V$;

(2) A is V-elliptic, i.e. $\exists\ \alpha > 0$ such that $A(v,v) \geq \alpha\|v\|_V^2 \quad \forall v \in V$;

(3) A is continuous, i.e. $\exists\, C \in \mathbb{R}$ such that $|A(v,w)| \le C\|v\|_V\|w\|_V$; and

(4) L is continuous, i.e. $\exists\, \Lambda \in \mathbb{R}$ such that $|L(v)| \le \Lambda\|v\|_V \quad \forall v \in V$.

Then there is a unique function $u \in V$ such that $A(u,v) = L(v) \quad \forall v \in V$, and the stability estimate $\|u\|_V \le \Lambda/\alpha$ holds.

Proof. The goal is to construct $u \in V$ solving the minimization problem $F(u) \le F(v)$ for all $v \in V$, which by the previous theorem is equivalent to the variational problem. The energy norm, $\|v\|^2 \equiv A(v,v)$, is equivalent to the norm of V, since by Condition 2 and 3,

$$\alpha\|v\|_V^2 \le A(v,v) = \|v\|^2 \le C\|v\|_V^2.$$

Let

$$\beta = inf_{v\in V} F(v). \tag{7.15}$$

Then $\beta \in \mathbf{R}$, since

$$F(v) = \frac{1}{2}\|v\|^2 - L(v) \ge \frac{1}{2}\|v\|^2 - \Lambda\|v\| \ge -\frac{\Lambda^2}{2}.$$

We want to find a solution to the minimization problem $\min_{v\in V} F(v)$. It is therefore natural to study a minimizing sequence v_i, such that

$$F(v_i) \to \beta = \inf_{v\in V} F(v). \tag{7.16}$$

The next step is to conclude that the v_i infact converge to a limit:

$$\begin{aligned}
\left\|\frac{v_i - v_j}{2}\right\|^2 &= \frac{1}{2}\|v_i\|^2 + \frac{1}{2}\|v_j\|^2 - \left\|\frac{v_i + v_j}{2}\right\|^2 \quad \text{(by the parallelogram law)} \\
&= \frac{1}{2}\|v_i\|^2 - L(v_i) + \frac{1}{2}\|v_j\|^2 - L(v_j) \\
&\quad - \left(\left\|\frac{v_i + v_j}{2}\right\|^2 - 2L(\frac{v_i + v_j}{2})\right) \\
&= F(v_i) + F(v_j) - 2F\left(\frac{v_i + v_j}{2}\right) \\
&\le F(v_i) + F(v_j) - 2\beta \quad \text{(by (7.15))} \\
&\to 0, \qquad \text{(by (7.16))}.
\end{aligned}$$

Hence $\{v_i\}$ is a Cauchy sequence in V and since V is a Hilbert space (in particular V is a complete space) we have $v_i \to u \in V$.

Finally $F(u) = \beta$, since

$$\begin{aligned} |F(v_i) - F(u)| &= |\frac{1}{2}(\|v_i\|^2 - \|u\|^2) - L(v_i - u)| \\ &= |\frac{1}{2}A(v_i - u, v_i + u) - L(v_i - u)| \\ &\leq (\frac{C}{2}\|v_i + u\|_V + \Lambda)\|v_i - u\|_V \\ &\to 0. \end{aligned}$$

Therefore there exists a unique (why?) function $u \in V$ such that $F(u) \leq F(v) \quad \forall v \in V$. To verify the stability estimate, take $v = u$ in (Var) and use the ellipcity (1) and continuity (3) to obtain

$$\alpha\|u\|_V^2 \leq A(u,u) = L(u) \leq \Lambda\|u\|_V$$

so that

$$\|u\|_V \leq \frac{\Lambda}{\alpha}.$$

The uniqueness of u can also be verified from the stability estimate. If u_1, u_2 are two solutions of the variational problem we have $A(u_1 - u_2, v) = 0$ for all $v \in V$. Therefore the stability estimate implies $\|u_1 - u_2\|_V = 0$, i.e. $u_1 = u_2$ and consequently the solution is unique. □

Example 7.9 Determine conditions for the functions k, r and $f : \Omega \to \mathbb{R}$ such that the assumptions in the Lax-Milgram theorem are satisfied for the following elliptic partial differential equation in $\Omega \subset \mathbf{R}^2$

$$\begin{aligned} -div(k\nabla u) + ru &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

Solution. This problem satisfies (Var) with

$$V = \{v \ : \ \int_\Omega (v^2(x) + |\nabla v(x)|^2)\ dx < \infty, \text{ and } v|_{\partial\Omega} = 0\},$$

$$
\begin{aligned}
A(u,v) &= \int_\Omega (k\nabla u\nabla v + ruv)\; dx,\\
L(v) &= \int_\Omega fv\; dx,\\
\|v\|_V^2 &= \int_\Omega (v^2(x) + |\nabla v|^2)\; dx.
\end{aligned}
$$

Consequently V is a Hilbert space and A is symmetric and continuous provided k and r are uniformly bounded.

The ellipticity follows by

$$
\begin{aligned}
A(v,v) &= \int_\Omega (k|\nabla v|^2 + rv^2)\; dx\\
&\geq \alpha\int_\Omega (v^2(x) + |\nabla v|^2)\; dx\\
&= \alpha\|v\|_{H^1}^2,
\end{aligned}
$$

provided $\alpha = \inf_{x\in\Omega}(k(x), r(x)) > 0$.

The continuity of A is a consequence of

$$
\begin{aligned}
A(v,w) &\leq \max(\|k\|_{L^\infty}, \|r\|_{L^\infty})\int_\Omega (|\nabla v||\nabla w| + |v||w|)dx\\
&\leq \max(\|k\|_{L^\infty}, \|r\|_{L^\infty})\|v\|_{H^1}\|w\|_{H^1},
\end{aligned}
$$

provided $\max(\|k\|_{L^\infty}, \|r\|_{L^\infty}) = C < \infty$.

Finally, the functional L is continuous, since

$$
|L(v)| \leq \|f\|_{L^2}\|v\|_{L^2} \leq \|f\|_{L^2}\|v\|_V,
$$

which means that we may take $\Lambda = \|f\|_{L^2}$ provided we assume that $f \in L^2(\Omega)$. Therefore the problem satisfies the Lax-Milgram theorem. $\square$

Example 7.10 Verify that the assumption of the Lax-Milgram theorem are satisfied for the following problem,

$$
\begin{aligned}
-\Delta u &= f \quad \text{in } \Omega,\\
u &= 0 \quad \text{on } \partial\Omega.
\end{aligned}
$$

Solution. This problem satisfies (Var) with

$$\begin{aligned} V = H_0^1 &= \{v \in H^1 \;:\; v|_{\partial\Omega} = 0\}, \\ H^1 &= \{v \;:\; \int_\Omega (v^2(x) + |\nabla v(x)|^2)\; dx < \infty\}, \end{aligned}$$

$$\begin{aligned} A(u,v) &= \int_\Omega \nabla u \nabla v \; dx, \\ L(v) &= \int_\Omega f v \; dx. \end{aligned}$$

To verify the V-ellipticity, we use the *Poincaré inequality*, i.e. there is a constant C such that

$$v \in H_0^1 \quad \Rightarrow \quad \int_\Omega v^2 \; dx \leq C \int_\Omega |\nabla u|^2 \; dx. \tag{7.17}$$

In one dimension and $\Omega = (0,1)$, the inequality (7.17) takes the form

$$\int_0^1 v^2(x) \; dx \leq \int_0^1 (v'(x))^2 \; dx, \tag{7.18}$$

provided $v(0) = 0$. Since

$$v(x) = v(0) + \int_0^x v'(s) \; ds = \int_0^x v'(s) \; ds,$$

and by Cauchy's inequality

$$\begin{aligned} v^2(x) = \left(\int_0^x v'(s) \; ds \right)^2 &\leq x \int_0^x v'(s)^2 \; ds \\ &\leq \int_0^1 v'(s)^2 \; ds \qquad \text{since } x \in (0,1). \end{aligned}$$

The V-ellipticity of A follows by (7.18) and

$$\begin{aligned} A(v,v) = \int_0^1 v'(x)^2 \; dx &= \frac{1}{2} \int_0^1 \left((v'(x))^2 \; dx + \frac{1}{2}(v'(x))^2 \right) \; dx \\ &\geq \frac{1}{2} \int_0^1 (v'(x)^2 + v(x)^2) \; dx \\ &= \frac{1}{2} \|v\|^2_{H_0^1} \qquad \forall v \in H_0^1. \end{aligned}$$

The other conditions can be proved similarly as in the previous example. Therefore this problem satisfies the Lax-Milgram theorem. $\square$

Chapter 8

Markov Chains, Duality and Dynamic Programming

8.1 Introduction

There are two main ideas in the arbitrage theory of pricing. One is that in complete markets, everyone should agree on a common price – any other price leads to an arbitrage opportunity. The other is that this price is the expected value of the cash flow with respect to some probability model – risk neutral pricing. In the simplest case, this probability model is a discrete Markov chain. This lecture describes how to compute probabilities and expected values for discrete Markov chain models. This is the main computational step in "risk neutral" option pricing.

The methods here compute the expected values by a time marching process that uses the transition matrix. Another evolution process allows us to compute probabilities. These evolution processes are related but not the same. The relation between the forward evolution for probabilities and the backward evolution for expected values is called *duality*. It is similar to the relation between a matrix and its transpose. The transpose of a matrix is sometimes called its dual.

The method of risk neutral arbitrage pricing extends to other more technical situations, but the main ideas are clear in the simple context of Markov chains. If the Markov chain model is replaced by a stochastic differential equation model, then the transition matrix is replaced by a partial differential operator – the "generator", and the matrix transpose is replaced by the "dual" of this generator. This is the subject of future lectures.

Many financial instruments allow the holder to make decisions along the way that effect the ultimate value of the instrument. American style options, loans that be repaid early, and convertible bonds are examples. To compute the value of such an instrument, we also seek the optimal decision strategy. *Dynamic programming* is a computational method that computes the value and decision strategy at the same time. It reduces the difficult "multiperiod decision problem" to a sequence of hopefully easier "single period" problems. It works backwards in time much as the expectation method does. The tree method commonly used to value American style stock options is an example of the general dynamic programming method.

8.2 Markov Chains

(This section assumes familiarity with basic probability theory using mathematicians' terminology. References on this include the probability books by G. C. Rota, W. Feller, Hoel and Stone, and B. V. Gnedenko.)

Many discrete time discrete state space stochastic models are stationary discrete Markov chains. Such a Markov chain is characterized by its state space, $\mathcal{S}$, and its transition matrix, P. We use the following notations:

- x, y, $\ldots$: possible states of the system, elements of $\mathcal{S}$.
- The possible times are $t = 0, 1, 2, \ldots$.
- $X(t)$: the (unknown) state of the system at time t. It is some element of $\mathcal{S}$.
- $u(x,t) = \mathbf{Pr}(X(t) = x)$. These probabilities satisfy an evolution equation moving forward in time. We use similar notation for conditional probabilities, for example, $u(x,t|X(0) = x_0) = \mathbf{Pr}(X(t) = x|X(0) = x_0)$.
- $p(x,y) = \mathbf{Pr}(x \to y) = \mathbf{Pr}(X(t+1) = y|X(t) = x)$. These "transition probabilities" are the elements of the transition matrix, P.

The transition probabilities have the properties:

$$0 \leq p(x,y) \leq 1 \quad \text{for all } x \in \mathcal{S} \text{ and } y \in \mathcal{S}. \tag{8.1}$$

and

$$\sum_{y \in \mathcal{S}} p(x,y) = 1 \quad \text{for all } x \in \mathcal{S}. \tag{8.2}$$

The first is because the $p(x,y)$ are probabilities, the second because the state x must go somewhere, possibly back to x. It is not true that

$$\text{(NOT ALWAYS TRUE)} \qquad \sum_{x \in \mathcal{S}} p(x,y) = 1 \;\; . \qquad \text{(NOT ALWAYS TRUE)}$$

The Markov property is that knowledge of the state at time t is all the information about the present and past relevant to predicting the future. That is:

$$\begin{aligned} \mathbf{Pr}(X(t+1) = y | X(t) = x_0, X(t-1) = x_1, \ldots) \\ = \mathbf{Pr}(X(t+1) = y | X(t) = x_0) \end{aligned} \tag{8.3}$$

no matter what extra history information $(X(t-1) = x_1, \ldots)$ we have. This may be thought of as a lack of long term memory. It may also be thought of as a completeness property of the model: the state space is rich enough to characterize the state of the system at time t completely.

To illustrate this point, consider the model

$$Z(t+1) = aZ(t) + bZ(t-1) + \xi(t) \;, \tag{8.4}$$

where the $\xi(t)$ are independent random variables. Models like this are used in "time series analysis". Here Z is a continuous variable instead a discrete variable to make the example simpler. If we say that the state at time t is $Z(t)$ then (8.4) is not a Markov chain. Clearly we do better at predicting $Z(t+1)$ if we know both $Z(t)$ and $Z(t-1)$ than if we know just $Z(t)$. If we say that the state at time t is the two dimensional vector

$$X(t) = \begin{pmatrix} Z(t) \\ Z(t-1) \end{pmatrix} \;\; ,$$

then

$$\begin{pmatrix} Z(t) \\ Z(t-1) \end{pmatrix} = \begin{pmatrix} a & b \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \xi(t) \\ 0 \end{pmatrix}$$

may be rewriten

$$X(t+1) = AX(t) + \begin{pmatrix} \xi(t) \\ 0 \end{pmatrix} \;.$$

Thus, $X(t)$ is a Markov chain. This trick of expressing lag models with multidimensional states is common in time series analysis.

The simpler of the evolutions, and the one less used in practice, is the forward evolution for the probabilities $u(x,t)$. Once we know the numbers $u(x,t)$ for all $x \in \mathcal{S}$ and a particular t, we can compute them for $t+1$. Proceding in this way, starting from the numbers $u(x,0)$ for all $x \in \mathcal{S}$, we can compute up to whatever T is desired. The evolution equation for the probabilities $u(x,t)$ is found using conditional probability:

$$
\begin{aligned}
u(x,t+1) &= \mathbf{Pr}(X(t+1)=x) \\
&= \sum_{y\in\mathcal{S}} \mathbf{Pr}(X(t+1)=x|X(t)=y)\cdot\mathbf{Pr}(X(t)=y) \\
u(x,t+1) &= \sum_{y\in\mathcal{S}} p(y,x)u(y,t) \ . \qquad (8.5)
\end{aligned}
$$

To express this in matrix form, we suppose that the state space, $\mathcal{S}$, is finite, and that the states have been numbered $x_1, \ldots, x_n$. The transition matrix, P, is $n\times n$ and has (i,j) entry $p_{ij} = p(x_i,x_j)$. We sometimes conflate i with x_i and write $p_{xy} = p(x,y)$; until you start programming the computer, there is no need to order the states. With this convention, (8.5) can be interpreted as vector–matrix multiplication if we define a *row* vector $\underline{u}(t)$ with components $(u_1(t),\ldots,u_n(t))$, where we have written $u_i(t)$ for $u(x_i,t)$. As long as ordering is unimportant, we could also write $u_x(t) = u(x,t)$. Now, (8.5) can be rewritten

$$\underline{u}(t+1) = \underline{u}(t)P \ . \qquad (8.6)$$

Since $\underline{u}$ is a row vector, the expression $P\underline{u}$ does not make sense because the dimensions of the matrices are incompatible for matrix multiplication. The convention of using a row vector for the probabilities and therefore putting the vector in the left of the matrix is common in applied probability. The relation (8.6) can be used repeatedly[1]

$$
\begin{gathered}
\underline{u}(1) = \underline{u}(0)P \text{ and } \underline{u}(2) = \underline{u}(1)P \\
\to \\
\underline{u}(2) = \left(\underline{u}(0)P\right)P = \underline{u}(0)\left(PP\right) = \underline{u}(0)P^2
\end{gathered}
$$

[1] The most important fact in linear algebra is that matrix multiplication is associative: $(AB)C = A(BC)$ for any three matrices of any size, including row or column vectors, as long as the multiplication is compatible.

to yield

$$\underline{u}(t) = \underline{u}(0)P^t \quad , \tag{8.7}$$

where P^t means P to the power t, not the transpose of P.

Actually, the Markov property is a bit stronger than (8.3). It applies not only to events determined by time $t+1$, but to any events determined in the future of t. For example, if A is the event $X(t+3) = x$ or y and $X(t+1) \neq X(t+4)$, then

$$\mathbf{Pr}(A \mid X(t) = z \text{ and } X(t-1) = w) = \mathbf{Pr}(A \mid X(t) = z) \; .$$

8.3 Expected Values

The more general and useful evolution equation is the backward evolution for expected values. In the simplest situation, suppose that $X(t)$ is a Markov chain, that the probability distribution $u(x,0) = \mathbf{Pr}(X(0) = x)$ is known, and that we want to evaluate $\mathbf{E}(V(X(T))$. We will call time $t = 0$ the present, time $t = T$ the payout time, and times $t = 1, \cdots, T-1$ intermediate times.

The backward evolution computed the desired expected value in terms of a collection of other conditional expected values, $f(x,t)$, where $x \in \mathcal{S}$ and t is an intermediate time. We start with the final time values $f(x,T) = V(x)$ for all $x \in \mathcal{S}$. We then compute the numbers $f(x, T-1)$ using the $f(x,t)$ and P. We continue in this way back to time $t = 0$.

The $f(x,t)$ are expected values of the payout, given knowledge of the state at a future intermediate time:

$$f(x,t) = \mathbf{E}\left[V(X(T)) | X(t) = x\right] \quad . \tag{8.8}$$

Recall our convention that time 0 is the present time, time $t > 0$ is in the future, but not as far in the future as the time, T, at which the payout is made. We may think of the $f(x,t)$ as possible expected values at the future intermediate time t. At time t we would know the value of $X(t)$. If that value were x, then the expected value of $V(X(T))$ would be $f(x,t)$.

Instead of computing $f(x,t)$ directly from the definition (8.8), we can compute it in terms of the $f(x,t+1)$ using the transition matrix. If the system is in state x at time t, then the probability for it to be at state y at the next time is $p(x \to y) = p(x,y)$. For expectation values, this implies

$$\begin{aligned} f(x,t) &= \mathbf{E}\left[f_T(X(T))|X(t)=x\right] \\ &= \sum_{y\in\mathcal{S}} \mathbf{E}\left[f_T(X(T))|X(t+1)=y\right]\cdot \mathbf{Pr}\left(X(t+1)=y \mid X(t)=x\right) \\ f(x,t) &= \sum_{y\in\mathcal{S}} f(y,t+1)p(x,y) \quad . \end{aligned} \tag{8.9}$$

It is clear from (8.8) that $f(x,T) = V(x)$; if we know the state at time T then we know the payout exactly. From these, we compute all the numbers $f(x,T-1)$ using (8.9) with $t = T-1$. Continuing like this, we eventually get to $t = 0$. We may know $X(0)$, the state of the system at the current time. For example, if $X(t)$ is the price of a stock at time t, then $X(0) = x_0$ is the current spot price. Then the desired expected value would be $f(x_0,0)$. Otherwise we can use

$$\begin{aligned} \mathbf{E}\left[V(X(T))\right] &= \sum_{x\in\mathcal{S}} \mathbf{E}\left[V(X(T))|X(0)=x\right]\cdot \mathbf{Pr}\left(X(0)=x\right) \\ &= \sum_{x\in\mathcal{S}} f(x,0)u(x,0) \quad . \end{aligned}$$

All the values on the bottom line should be known.

Another remark on the interpretation of (8.9) will be helpful. Suppose we are at state x at time t and wish to know the expected value of $V(X(T))$. In one time step, starting from state x, we could go to state y at time $t+1$ with probability[2] $p(x,y)$. The right side of (8.9) is the average over the possible y values, using probability $p(x,y)$. The quantities being averaged, $f(y,t+1)$ are themselves expected values of $V(X(T))$. Thus, we can read (8.9) as saying that the expected value is the expected value of the expected values at the next time. A simple model for this situation is that we toss a coin. With probability p we get payout U and with probability $1-p$ we get payout V. Let us suppose that both U and V are random with expected values $f_U = \mathbf{E}(U)$ and $f_V = \mathbf{E}(V)$. The overall expected payout is $p \cdot f_u + (1-p)\cdot f_V$. The Markov chain situation is like this. We are at a state x at time t. We first choose state $y \in \mathcal{S}$ with probability $p(x,y)$. For each y at time $t+1$ there is a payout probability, U_y, whose probability distribution depends on y, $t+1$,

[2]Here we should think of y as the variable and x as a parameter.

V, and the Markov chain. The overall expected payout is the average of the expected values of the U_y, which is what (8.9) says.

As with the probability evolution equation (8.5), the equation for the evolution of the expectation values (8.9) can be written in matrix form. The difference from the probability evolution equation is that here we arrange the numbers $f_j = f(x_j, t)$ into a *column* vector, $\underline{f(t)}$. The evolution equation for the expectation values is then written in matrix form as

$$\underline{f}(t) = P\underline{f}(t+1) \quad . \tag{8.10}$$

This time, the vector goes on the right. If apply (8.10) repeatedly, we get, in place of (8.7),

$$\underline{f}(t) = P^{T-t}\underline{f}(T) \quad . \tag{8.11}$$

There are several useful variations on this theme. For example, suppose that we have a running payout rather than a final time payout. Call this payout $g(x,t)$. If $X(t) = x$ then $g(x,t)$ is added to the total payout that accumulates over time from $t = 0$ to $t = T$. We want to compute

$$\mathbf{E}\left[\sum_{t=0}^{T} g(X(t),t)\right] \quad .$$

As before, we find this by computing more specific expected values:

$$f(x,t) = \mathbf{E}\left[\sum_{t'=t}^{T} g(X(t'),t') | X(t) = x\right] \quad .$$

These numbers are related through a generalization of (8.9) that takes into account the known contribution to the sum from the state at time t:

$$f(x,t) = \sum_{y \in \mathcal{S}} f(y,t+1)p(x,y) + g(x,t) \quad .$$

The "initial condition", given at the final time, is

$$f(x,T) = g(x,T) \quad .$$

This includes the previous case, we take $g(x,T) = f_T(x)$ and $g(x,t) = 0$ for $t < T$.

As a final example, consider a path dependent discounting. Suppose for a state x at time t there is a discount factor $r(x,t)$ in the range $0 \leq r(x,t) \leq 1$. A cash flow worth f at time $t+1$ will be worth $r(x,t)f$ at time t if $X(t) = x$. We want the discounted value at time $t = 0$ at state $X(0) = x$ of a final time payout worth $f_T(X(T))$ at time T. Define $f(x,t)$ to be the value at time t of this payout, given that $X(t) = x$. If $X(t) = x$ then the time $t+1$ expected discounted (to time $t+1$) value is

$$\sum_{y \in \mathcal{S}} f(y, t+1) p(x,y) \quad .$$

This must be discounted to get the time t value, the result being

$$f(x,t) = r(x,t) \sum_{y \in \mathcal{S}} f(y, t+1) p(x,y) \quad .$$

8.4 Duality and Qualitative Properties

The forward evolution equation (8.5) and the backward equation (8.9) are connected through a duality relation. For any time t, we compute (8.8) as

$$\begin{aligned} \mathbf{E}\left[V(X(T))\right] &= \sum_{x \in \mathcal{S}} \mathbf{E}\left[V(X(T)) | X(t) = x\right] \cdot \mathbf{Pr}(X(t) = x) \\ &= \sum_{x \in \mathcal{S}} f(x,t) u(x,t) \quad . \end{aligned} \tag{8.12}$$

For now, the main point is that the sum on the bottom line does not depend on t. Given the constancy of this sum and the u evolution equation (8.5), we can give another derivation of the f evolution equation (8.9). Start with

$$\sum_{x \in \mathcal{S}} f(x, t+1) u(x, t+1) = \sum_{y \in \mathcal{S}} f(y,t) u(y,t) \quad .$$

Then use (8.5) on the left side and rearrange the sum:

$$\sum_{y \in \mathcal{S}} \left(\sum_{x \in \mathcal{S}} f(x, t+1) p(y,x) \right) u(y,t) = \sum_{y \in \mathcal{S}} f(y,t) u(y,t) \quad .$$

Now, if this is going to be true for any $u(y,t)$, the coefficients of $u(y,t)$ on the left and right sides must be equal for each y. This gives (8.9). Similarly, it is possible to derive (8.5) from (8.9) and the constancy of the expected value.

The evolution equations (8.5) and (8.9) have some qualitative properties in common. The main one being that they preserve positivity. If $u(x,t) \geq 0$ for all $x \in \mathcal{S}$, then $u(x,t+1) \geq 0$ for all $x \in \mathcal{S}$ also. Likewise, if $f(x,t+1) \geq 0$ for all x, then $f(x,t) \geq 0$ for all x. These properties are simple consequences of (8.5) and (8.9) and the positivity of the $p(x,y)$. Positivity preservation does not work in reverse. It is possible, for example, that $f(x,t+1) < 0$ for some x even though $f(x,t) \geq 0$ for all x.

The probability evolution equation (8.5) has a conservation law not shared by (8.9). It is

$$\sum_{x \in \mathcal{S}} u(x,t) = \mathit{const}\,. \tag{8.13}$$

independent of t. This is natural if u is a probability distribution, so that the constant is 1. The expected value evolution equation (8.9) has a *maximum principle*

$$\max_{x \in \mathcal{S}} f(x,t) \leq \max_{x \in \mathcal{S}} f(x,t+1) \quad . \tag{8.14}$$

This is a natural consequence of the interpretation of f as an expectation value. The probabilities, $u(x,t)$ need not satisfy a maximum principle either forward of backward in time.

This duality relation has is particularly transparent in matrix terms. The formula (8.8) is expressed explicitly in terms of the probabilities at time t as

$$\sum_{x \in \mathcal{S}} f(x,T)u(x,T) \quad ,$$

which has the matrix form

$$\underline{u}(t)\underline{f}(T) \quad .$$

Written in this order, the matrix multiplication is compatible; the other order, $\underline{f}(T)\underline{u}(T)$, would represent an $n \times n$ matrix instead of a single number. In view of (8.7), we may rewrite this as

$$\underline{u}(0)P^t\underline{f}(T) \quad .$$

Because matrix multiplication is associative, this may be rewritten

$$\left[\underline{u}(0)P^t\right] \cdot \left[P^{T-t}\underline{f}(T)\right] \tag{8.15}$$

for any t. This is the same as saying that $\underline{u}(t)\underline{f}(T-t)$ is independent of t, as we already saw.

In linear algebra and functional analysis, "adjoint" or "dual" is a fancy generalization of the transpose operation of matrices. People who don't like to think of putting the vector to the left of the matrix think of $\underline{u}P$ as multiplication of (the transpose of) $\underline{u}$, on the right, by the transpose (or adjoint or dual) of P. In other words, we can do enough evolution to compute an expected value either using P its dual (or adjoint or transpose). This is the origin of the term "duality" in this context.

8.5 Dynamic Programming

Dynamic programming is a method for valuing American style options and other financial instruments that allow the holder to make decisions that effect the ultimate payout. The idea is to define the appropriate value function, $f(x,t)$, that satisfies a nonlinear version of the backwards evolution equation (8.9). In the real world, dynamic programming is used to determine "optimal" trading strategies for traders trying to take or unload a big position without moving the market, to find cost efficient hedging strategies when trading costs or other market frictions are significant, and for many other purposes. Its main drawback stems from the necessity of computing the cost to go function (see below) for every state $x \in \mathcal{S}$. For complex models, the state space may be too large for this to be practical. That's when things really get interesting.

I will explain the idea in a simple but somewhat abstract situation. As in the previous section, it is possible to use these ideas to treat other related problems. We have a Markov chain as before, but now the transition probabilities depend on a "control parameter", ξ. That is

$$p(x,y,\xi) = \mathbf{Pr}\left(X(t+1) = y | X(t) = x, \xi\right) \quad .$$

In the "stochastic control problem", we are allowed to choose the control parameter at time t, $\xi(t)$, knowing the value of $X(t)$ but not any more about the future than the transition probabilities. Because the system is a Markov chain, knowledge of earlier values, $X(t-1), \ldots$, will not help predict or control the future. Choosing ξ as a function of $X(t)$ and t is called "feedback control" or a "decision strategy". The point here is that the optimal control policy is a feedback control. That is, instead of trying to choose a whole control trajectory, $\xi(t)$ for $t = 0, 1, \ldots, T$, we instead try to choose the feedback functions $\xi(X(t), t)$. We will write $\xi(X, t)$ for such a decision strategy.

Any given strategy has an expected payout, which we write

$$\mathbf{E}_\xi \left[V(X(T)) \right] \quad .$$

Our object is to compute the value of the financial instrument under the optimal decision strategy:

$$\max_\xi \mathbf{E}_\xi \left[V(X(T)) \right] \quad , \tag{8.16}$$

and the optimal strategy that achieves this.

The appropriate collection of values for this is the "cost to go" function

$$f(x, t) = \max_\xi \mathbf{E}_\xi \left[V(X(T)) | X(t) = x \right] \quad . \tag{8.17}$$

As before, we have "initial data" $f(x, T) = V(x)$. We need to compute the values $f(x, t)$ in terms of already computed values $f(x, t+1)$. For this, we suppose that the optimal decision strategy at time t is not yet known but those at later times are already computed. If we use control variable $\xi(t)$ at time t, and the optimal control thereafter, we get payout depending on the state at time $t + 1$:

$$\mathbf{E} \left[f(X(t+1), t+1) | X(t) = x, \xi(t) \right] = \sum_{y \in \mathcal{S}} f(y, t+1) p(x, y, \xi(t)) \quad .$$

Maximizing this expected payout over $\xi(t)$ gives the optimal expected payout at time t:

$$f(x, t) = \max_{\xi(t)} \sum_{y \in \mathcal{S}} f(y, t+1) p(x, y, \xi(t)) \quad . \tag{8.18}$$

This is the principle of dynamic programming. We replace the "multiperiod optimization problem" (8.17) with a sequence of hopefully simpler "single period" optimization problems (8.18) for the cost to go function.

8.6 Examples and Exercises

1. A stationary Markov chain has three states, called A, B, and C. The probability of going from A to B in one step is .6. The probability of staying at A is .4. The probability of going from B to A is .3. The probability of staying at B is .2, and the probability of going to C is .5. From state C, the probability of going to B is .8 and the probability of going to A is zero. The payout for state A is 1, for state B is 4, and for state C is 9.

a. Compute the probabilities that the system will be in state A, B, or C after two steps, starting from state A. Use these three numbers to compute the expected payout after two steps starting from state A.

b. Compute the expected payouts in one step starting from state A and from state B. These are $f(A,1)$ and $f(B,1)$ respectively.

c. See that the appropriate average of $f(A,1)$ and $f(B,1)$ agrees with the answer from part a.

2. Suppose a stock price is a stationary Markov chain with the following transition probabilities. In one step, the stock goes from S to uS with probability p and from S to dS with probability $q = 1 - p$. We generally suppose that u (the uptick) is slightly bigger than one while d (the downtick) as a bit smaller. Show that the method for computing the expected payout is exactly the binomial tree method for valuing European style options.

3. Formulate the American style option valuation problem as an optimal decision problem. Choosing the early exercise time is the same as deciding on each day whether to exercise or not. Show that the dynamic probramming algorithm discussed above is the binomial tree method for Amercian style options. The optimization problem (8.18) reduces to taking the max between the computed f and the intrinsic value.

4. This is the simplest example of the "linear quadratic gaussian" (LQG) paradigm in optimal control that has become the backbone of traditional control engineering. Here $X(t)$ is a real number. The transitions are given by

$$X(t+1) = aX(t) + \sigma G(t) + \xi(t) \;, \qquad (8.19)$$

where $G(t)$ is a standard normal random variable and the $G(t)$ for different t values are independent. We want to minimize the quantity

$$C = \sum_{t=1}^{T} X(t)^2 + \mu \sum_{t=0}^{T-1} \xi(t)^2 \tag{8.20}$$

We want to find a chioce of the control, ξ, that minimizes $\mathbf{E}(C)$. Note that the dynamics (8.19) are linear, the noise is gaussian, and the cost function (8.20) is quadratic. Define the cost to go function $f(x, t)$ to be the cost incurred starting at x at time t ignoring the costs that are incurred at earlier times. Start by computing $f(x, T-1)$ explicitly by minimizing over the single variable $\xi(T-1)$. Note that the optimal $\xi(T-1)$ is a linear function of $X(T-1)$. Using this information, compute $f(x, T-2)$ by optimizing over $\xi(T-2)$, and so on. The LQG model in control engineering justifies linear feedback control in much the same way the gaussian error model and maximum likelihood justifies least squares estimation in statistics.

Chapter 9

Optimal Control

The purpose of Optimal Control is to influence the behavior of a dynamical system in order to achieve a desired goal. Optimal control has a huge variety of applications, such as aerospace, aeronautics, chemical plants, mechanical systems, finance and economics. In this chapter we consider dynamical systems whose evolution is determined by ordinary stochastic differential equations, although the derived principles are still valid in more general situations.

To give some intuition on the subject and to introduce some basic concepts let us consider a hydro-power generator in a river. Suppose that we are the owners of such a generator, and that our goal is to maximise our profit by selling electricity in some local electricity market. This market will offer us buying prices at different hours, so one decision we have to make is when and how much electricity to generate. To make this decision may not be a trivial task, since besides economic considerations, we also have to meet technical constraints. For instance, the power generated is related to the amount of water in the reservoir, the turbined flow and other variables. Moreover, if we want a plan for a period longer than just a few days the water inflow to the lake may not be precisely known, making the problem stochastic.

We can now state our problem then in optimal control terms as the maximization of an *objective function*, the expected profit from selling electricity power during a given period, with respect to control functions, like the hourly turbined flow. Observe that the turbined flow is positive and smaller than a given maximum value, so it is natural to have a set of *feasible controls*, namely the set of those controls we can use in practice. In addition, our dynamical system evolves according to a given law, also called *the dynamics*, which here comes from a mass balance in the dam's lake. This law tells us how the *state variable*, the amount of water in the lake, evolves with time according to the control we give. Since the volume in the lake cannot be negative, there exist additional constraints, known as *state constraints*, that have to be fulfilled in the optimal control problem.

After introducing the formulation of an optimal control problem the next step is to find its solution. As we shall see, the optimal control is closely related with the solution of a nonlinear partial differential equation, known as the Hamilton Jacobi equation. To derive the Hamilton Jacobi equation we shall use the dynamic programming principle, which relates the solution of a given optimal control problem with solutions to simpler problems.

9.1 An Optimal Portfolio

Example 9.1 *Assume that the value of a portfolio, $X(t)$, consists of risky stocks, $S(t) = \alpha(t)X(t)$, and riskless bonds, $B(t) = (1-\alpha(t))X(t)$, where $\alpha(t) \in [0,1]$ and*

$$\begin{aligned} dS &= aSdt + cSdW, && (9.1)\\ dB &= bBdt, && (9.2) \end{aligned}$$

with $0 \le b < a$. Define for a given function g the cost function

$$C_{t,x}(\alpha) = E[g(X(T))|X(t) = x].$$

Then our goal is to determine the Markov control function $\alpha(t, X(t))$, with $\alpha : [0,T] \times \mathbb{R} \to [0,1]$ that maximizes the cost function. The solution will be based on the function

$$u(t,x) \equiv \max_{\alpha} C_{t,x}(\alpha),$$

and we will show that $u(t,x)$ satisfies the following *Hamilton-Jacobi* equation,

$$\begin{aligned} &u_t + \max_{\alpha\in[0,1]} \left\{ (a\alpha + b(1-\alpha))xu_x + \frac{c^2\alpha^2}{2}x^2u_{xx} \right\} = 0, && (9.3)\\ &u(T,x) = g(x), \end{aligned}$$

i.e.

$$u_t + H(x, u_x, u_{xx}) = 0$$

for

$$H(x,p,w) \equiv \max_{v\in[0,1]}(av + b(1-v)xp + \frac{c^2v^2}{2}x^2w).$$

Example 9.2 Assume that $u_{xx} < 0$ in the equation (9.3). Determine the optimal control function α^*.

Solution. By differentiating $f(\alpha) = (a\alpha + b(1-\alpha))xu_x + \frac{c^2\alpha^2}{2}x^2u_{xx}$ in (9.3) with respect to α and using $df/d\alpha = 0$, we obtain

$$\hat{\alpha} = -\frac{(a-b)u_x}{c^2xu_{xx}}.$$

Then the optimal control α^* is given by

$$\alpha^* = \begin{cases} 0, & if\ \hat{\alpha} < 0 \\ \hat{\alpha}, & if\ \hat{\alpha} \in [0,1] \\ 1 & if\ 1 < \hat{\alpha} \end{cases}$$

The optimal value yields in (9.3) the Hamilton-Jacobi equation

$$u_t + H(x, u_x, u_{xx}) = 0,$$

where

$$H(x,u_x,u_{xx}) = \begin{cases} bxu_x, \text{ if } \hat{\alpha} < 0 \\ bxu_x - \dfrac{(a-b)^2u_x^2}{2c^2u_{xx}}, \text{ if } \hat{\alpha} \in [0,1] \\ axu_x + \dfrac{c^2x^2u_{xx}}{2} \text{ if } 1 < \hat{\alpha} \end{cases} \tag{9.4}$$

□

Example 9.3 What is the optimal control function $\alpha = \alpha^*$ for $g(x) = x^r, 0 < r < 1$?

Solution. We have $dX = d(\alpha X + (1-\alpha)X) = dS + dB = (aS + bB)dt + cSdW = (a\alpha X + b(1-\alpha)X)dt + c\alpha X dW$, so that the Itô formula yields

$$\begin{aligned} dg(X) &= dX^r = rX^{r-1}dX + \frac{r(r-1)}{2}X^{r-2}(dX)^2 \\ &= rX^r(a\alpha + b(1-\alpha))dt + rX^r\alpha c dW + \frac{1}{2}\alpha^2c^2r(r-1)X^r dt. \end{aligned}$$

Taking the expectation value in the above,

$$E[X^r(T)] = X^r(0) + E\left[\int_0^T rX^r\left(a\alpha + b(1-\alpha) + \frac{1}{2}\alpha^2c^2(r-1)\right)dt\right].$$

Finally, perturb the above equation with respect to $\epsilon \in \mathbb{R}_+$ provided $\alpha = \alpha^* + \epsilon v$ for some feasible function v, that is $\alpha^* + \epsilon v \in [0,1]$ for sufficiently small ϵ. Then the optimal control, α^*, should satisfy $E[X^r_{\alpha^*+\epsilon v}(T)] - E[X^r_{\alpha^*}(T)] \geq 0\ \forall v$. If we make the assumption $\alpha^* \in (0,1)$, then we obtain

$$E[\int_0^T rX^r v(a - b + \alpha^* c^2(r-1))dt] = 0,\ \forall v$$

which implies

$$\alpha^* = \frac{a-b}{c^2(1-r)}.$$

$\square$

Exercise 9.4 What is the optimal control in (9.3) for $g(x) = \log x$?

9.2 Control of SDE

In this section we study optimal control of the solution $X(t)$ to the stochastic differential equation

$$\begin{cases} dX_i &= a_i(X(s), \alpha(s, X(s)))dt + b_{ij}(X(s), \alpha(s, X(s)))dW_j, \quad t < s < T \\ X(t) &= x \end{cases} \tag{9.5}$$

where T is a fixed terminal time and $x \in \mathbb{R}^n$ is a given initial point. Assume that $a_i, b_{ij} : \mathbb{R}^n \times A \to \mathbb{R}$ are smooth bounded functions, where A is a given compact subset of $\mathbb{R}^m$. The function $\alpha : [0,T] \times \mathbb{R}^n \to A$ is a *control* and let $\mathcal{A}$ be the set of admissible Markov control functions $t \to \alpha(t, X(t))$. The Markov control functions use the current value $X(s)$ to affect the dynamics of X by adjusting the drift and the diffusion coefficients. Let us for these admissible controls $\alpha \in \mathcal{A}$ define the *cost*

$$C_{t,x}(\alpha) = E[\int_t^T h(X(s), \alpha(s))ds + g(X(T))]$$

where X solves the stochastic differential equation (9.5) with control α and

$$h : \mathbb{R}^n \times A \to \mathbb{R}, \quad g : \mathbb{R}^n \to \mathbb{R}$$

are given smooth bounded functions. We call h the *running cost* and g the *terminal cost*. Our goal is to find an optimal control α^* which minimizes the expected cost, $C_{t,x}(\alpha)$.

Let us define the value function

$$u(t,x) \equiv \inf_{\alpha \in \mathcal{A}} C_{t,x}(\alpha). \tag{9.6}$$

The plan is to show that u solves a certain Hamilton-Jacobi equation and that the optimal control can be reconstructed from u. We first assume for simplicity that the optimal control is attained, i.e

$$u(t,x) = \min_{\alpha \in \mathcal{A}} C_{t,x}(\alpha) = C_{t,x}(\alpha^*).$$

The generalization of the proofs without this assumption is discussed in Exercise 9.7.

9.3 Dynamic Programming and Hamilton-Jacobi Equations

Lemma 9.5 *Assume that assumptions in section 9.1 hold. Then, the function u satisfies, for all $\delta > 0$, the dynamic programming relation*

$$u(t,x) = \min_{\alpha:[t,t+\delta]\to A} E[\int_t^{t+\delta} h(X(s), \alpha(s, X(s)))ds + u(t+\delta, X(t+\delta))]. \tag{9.7}$$

Proof. The proof has two steps: to use the optimal control to verify

$$u(t,x) \geq \min_{\alpha \in \mathcal{A}} E[\int_t^{t+\delta} h(X(s), \alpha(s))ds + u(t+\delta, X(t+\delta))],$$

and then to show that an arbitrary control yields

$$u(t,x) \leq \min_{\alpha \in \mathcal{A}} E[\int_t^{t+\delta} h(X(s), \alpha(s))ds + u(t+\delta, X(t+\delta))],$$

which together imply Lemma 9.5.

Step 1: Choose the optimal control α^*, from t to T, to obtain

$$\begin{aligned}
u(t,x) &= \min_{\alpha \in \mathcal{A}} E\left[\int_t^T h(X(s), \alpha(s, X(s)))ds + g(X(T))\right] \\
&= E[\int_t^{t+\delta} h(X(s), \alpha^*(s))ds] + E[\int_{t+\delta}^T h(X(s), \alpha^*(s))ds + g(X(T))] \\
&= E[\int_t^{t+\delta} h(X(s), \alpha^*(s))ds] \\
&\quad + E\left[E[\int_{t+\delta}^T h(X(s), \alpha^*(s))ds + g(X(T))|\ X(t+\delta)]\right] \\
&\geq E[\int_t^{t+\delta} h(X(s), \alpha^*(s))ds] + E[u(X(t+\delta), t+\delta)] \\
&\geq \min_{\alpha \in \mathcal{A}} E\left[\int_t^{t+\delta} h(X(s), \alpha(s, X(s))ds + u(X(t+\delta), t+\delta)\right].
\end{aligned}$$

Step 2: Choose the control α^+ to be arbitrary from t to $t+\delta$ and then, given the value $X(t+\delta)$, choose the optimal α^* from $t+\delta$ to T. Denote this control by $\alpha' = (\alpha^+, \alpha^*)$. Definition (9.6) shows

$$
\begin{aligned}
u(t,x) &\le C_{t,x}(\alpha') \\
&= E[\int_t^T h(X(s),\alpha'(s))ds + g(X(T))] \\
&= E[\int_t^{t+\delta} h(X(s),\alpha^+(s))ds] + E[\int_{t+\delta}^T h(X(s),\alpha^*(s))ds + g(X(T))] \\
&= E[\int_t^{t+\delta} h(X(s),\alpha^+(s))ds] \\
&\quad + E\left[E[\int_{t+\delta}^T h(X(s),\alpha^*(s))ds + g(X(T))|\ X(t+\delta)]\right] \\
&= E[\int_t^{t+\delta} h(X(s),\alpha^+(s))ds] + E[u(X(t+\delta),t+\delta)].
\end{aligned}
$$

Taking the minimum over all controls yields

$$
u(t,x) \le \min_{\alpha^+\in\mathcal{A}} E\left[\int_t^{t+\delta} h(X(s),\alpha^+(s))ds + u(X(t+\delta),t+\delta)\right].
$$

$\square$

Theorem 9.6 *Assume that X solves (9.5) with a Markov control function α and that the function u defined by (9.6) is bounded and smooth. Then u satisfies the Hamilton-Jacobi equation*

$$
\begin{aligned}
&u_t + H(t,x,Du,D^2u) = 0, \\
&u(T,x) = g(x),
\end{aligned}
$$

with the Hamiltonian function

$$
H(t,x,Du,D^2u) \equiv \min_{\alpha\in A}\left[a_i(x,\alpha)\partial_{x_i}u(t,x) + \frac{b_{ik}(x,\alpha)b_{jk}(x,\alpha)}{2}\partial_{x_ix_j}u(t,x) + h(x,\alpha)\right]
$$

Proof. The proof has two steps: to show that the optimal control $\alpha = \alpha^*$ yields

$$
u_t + a_i^*\partial_{x_i}u + \frac{b_{ik}^*b_{jk}^*}{2}\partial_{x_ix_j}u + h^* = 0, \tag{9.8}
$$

where $a^*(x) = a(x, \alpha^*(t,x)), b^*(x) = b(x, \alpha^*(t,x))$ and $h^*(t,x) = h(t,x,\alpha^*(t,x))$, and then that an arbitrary control α^+ implies

$$u_t + a_i^+ \partial_{x_i} u + \frac{b_{ik}^+ b_{jk}^+}{2} \partial_{x_i x_j} u + h^+ \geq 0, \tag{9.9}$$

where $a^+(x) = a(x, \alpha^+(t,x)), b^+(x) = b(x, \alpha^+(t,x))$ and $h^+(t,x) = h(t,x,\alpha^+(t,x))$. The two equations (9.8) and (9.9) together imply Theorem 9.6.

Step 1 : Choose $\alpha = \alpha^*$ to be the optimal control in (9.5). Then by the dynamic programming principle of Lemma 9.6

$$u(X(t),t) = E[\int_t^{t+\delta} h(X(s), \alpha^*(s, X(s)))ds + u(X(t+\delta), t+\delta)],$$

so that Itô 's formula implies

$$\begin{aligned}
-h(t,x,\alpha^*(t,x))dt &= E[du(X(t),t)|\ X(t) = x] \qquad (9.10)\\
&= (u_t + a_i^* \partial_{x_i} u + \frac{b_{ik}^* b_{jk}^*}{2} \partial_{x_i x_j} u)(t,x)dt.
\end{aligned}$$

Definition (9.6) shows

$$u(T,x) = g(x),$$

which together with (9.10) prove (9.8).

Step 2 : Choose the control function in (9.5) to be arbitrary from time t to $t+\delta$ and denote this choice by $\alpha = \alpha^+$. The function u then satisfies by Lemma 9.6

$$u(t,x) \ \leq \ E[\int_t^{t+\delta} h(X(s), \alpha^+(s))ds] + E[u(X(t+\delta), t+\delta)].$$

Hence $E[du] \geq -h(x, \alpha^+)dt$. We know that for any given α^+, by Itô 's formula,

$$E[du(t,X(t))] = E\left[u_t + a_i^+ \partial_{x_i} u + \frac{b_{ik}^+ b_{jk}^+}{2} \partial_{x_i x_j} u\right] dt.$$

Therefore, for any control α^+,

$$u_t + a_i^+ \partial_{x_i} u + \frac{b_{ik}^+ b_{jk}^+}{2} \partial_{x_i x_j} u + h(x, \alpha^+) \geq 0,$$

which proves (9.9) $\square$

Exercise 9.7 Use a minimizing sequence α_i of controls, satisfying

$$u(t,x) = \lim_{i\to\infty} C_{t,x}(\alpha_i),$$

to prove Lemma 9.6 and Theorem 9.6 without the assumption that the minimum control is attained.

Exercise 9.8 Let $\mathcal{A}^+$ be the set of all adapted controls $\{\alpha : [0,T]\times\mathcal{C}[0,T] \to A\}$ where $\alpha(s,X)$ may depend on $\{X(\tau) : \tau \le s\}$. Show that the minimum over all adapted controls in $\mathcal{A}^+$ is in fact the same as the minimum over all Markov controls, i.e.

$$\inf_{\alpha\in\mathcal{A}^+} C_{t,x}(\alpha) = \inf_{\alpha\in\mathcal{A}} C_{t,x}(\alpha),$$

e.g. by proving the dynamic programming relation (9.7) for adapted controls and motivate why this is sufficient.

9.4 Relation of Hamilton-Jacobi Equations and Conservation Laws

In this section we will analyze qualitative behavior of Hamilton-Jacobi equations, in particular we will study the limit corresponding to vanishing noise in control of stochastic differential equations. The study uses the relation between the Hamilton-Jacobi equation for $V : [0,T]\times\mathbb{R}\to\mathbb{R}$

$$V_t + H(V_x) = 0, \quad V(0,x) = V_0(x), \qquad (H-J)$$

and the conservation law for $U : [0,T]\times\mathbb{R}\to\mathbb{R}$

$$U_t + H(U)_x = 0, \quad U(0,x) = U_0(x). \qquad (C-L)$$

Observe that the substitution $V(t,x) = \int_{-\infty}^x U(t,y)dy$, so that $U = V_x$, and integration in x from $-\infty$ to x in (C-L) shows

$$V_t + H(V_x) = H(U(t,-\infty)). \tag{9.11}$$

Combined with the assumptions $U(t,x)\to 0$ as $|x|\to\infty$ and $H(0) = 0$ we conclude that V solves (H-J), if U solves (C-L).

Figure 9.1: Left: Initial condition. Right: Colliding characteristics and a shock.

The next step is to understand the nature of the solutions of (C-L). Consider the special Burger's conservation law

$$0 = U_t + U\ U_x = U_t + (\frac{U^2}{2})_x, \quad U(0,x) = U_0(x). \tag{9.12}$$

Let us define a *characteristic path* $X : [0,T] \times \mathbb{R} \to \mathbb{R}$ by

$$\frac{dX}{dt}(t) = U(t, X(t)),\ X(0) = x_0. \tag{9.13}$$

Thus, if $\psi(t) \equiv U(t, X(t))$ then $\frac{d\psi}{dt}(t) = 0$ by virtue of (9.12). This means that the value of U is constant along a characteristic path. If the characteristics do not collide into each other we may expect to find a solution using the initial data $U_0(x)$ and the set of characteristics. Unfortunately, this is not what happens in general, and collisions between characteristics do exist and give birth to discontinuities known as shocks. For example, this is the case when $U_0(x) = -\arctan(x)$ and $t \geq 1$.

Exercise 9.9 Show that $w(t) = U_x(X(t), t)$ satisfies $w(t) = w(0)/(1 + w(0)t)$, $t < 1$, for Burger's equation (9.12) with initial data $U(x,0) = -\arctan(x)$. Hence, $w(1) = \infty$, for $X(0) = 0$.

Since the method of characteristics does not work globally we have to find an alternative way to explain what happens with the solution $U(t,x)$ near a shock. It is not enough with the concept of strong or classical solution, since the solution $U(t,x)$ is not differentiable in general. For this purpose, we define the notion of weak solution. Let V be the set of test functions $\{\varphi : (0,+\infty) \times \mathbb{R} \to \mathbb{R}\}$ which are differentiable and take the value zero outside some compact set. Then an integrable function U is a weak solution of (9.12) if it satisfies

$$\int_0^{+\infty} \int_{-\infty}^{+\infty} \left(U(t,x)\varphi_t(t,x) + \frac{U^2(t,x)}{2}\varphi_x(t,x) \right) dx\ dt = 0,\ \forall \varphi \in V \tag{9.14}$$

and

$$\int_{-\infty}^{+\infty} |U(t,x) - U_0(x)| dx \to 0, \quad as\ t \to 0 \tag{9.15}$$

Figure 9.2: Shock velocity and Rankine Hugoniot condition

Example 9.10 The shock wave

$$U(t,x) = \begin{cases} 1 & x < \frac{t}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

is a weak solution satisfying (9.14) and (9.15). Observe that for $s \equiv 1/2$

$$\partial_t \int_a^b U \ dx = \frac{U^2(t,a) - U^2(t,b)}{2} = -\left[\frac{U^2}{2}\right],$$

and

$$\partial_t \int_a^b U \ dx = \partial_t[(s\ t - a)u_-] + (b - s\ t)u_+] = -s(U_+ - U_-),$$

where

$$[w(x_0)] \equiv w_+(x_0) - w_-(x_0) \equiv \lim_{y \to 0+} w(x_0 + y) - w(x_0 - y)$$

is the jump at the point x_0. Consequently, the speed s of a shock can be determined by the so called *Rankine Hugoniot* condition

$$s[U] = \left[\frac{U^2}{2}\right]. \tag{9.16}$$

Exercise 9.11 Verify that the shock wave solution

$$U_I(t,x) = \begin{cases} 0 & x > -\frac{t}{2}, \\ -1 & \text{otherwise} \end{cases}$$

and the rarefaction wave solution

$$U_{II}(t,x) = \begin{cases} 0 & x \geq 0, \\ \frac{x}{t} & -t < x < 0, \\ -1 & \text{otherwise} \end{cases}$$

are both weak solutions of $U_t + U\ U_x = 0$ with the same initial condition.

Figure 9.3: $U_I(t,x)$

Figure 9.4: $U_{II}(t,x)$

The last exercise shows that we pay a price to work with weak solutions: the lack of uniqueness. Therefore, we need some additional physical information to determine a unique weak solution. This leads us to the concept of *viscosity limit* or *viscosity solution*: briefly, it says that the weak solution U we seek is the limit $U = \lim_{\epsilon\to 0+} U^\epsilon$ of the solution of the regularized equation

$$U_t^\epsilon + U^\epsilon\ U_x^\epsilon = \epsilon U_{xx}^\epsilon, \quad \epsilon > 0. \tag{9.17}$$

This regularized equation has continuous and smooth solutions for $\epsilon > 0$. With reference to the previous example, the weak solution U_{II} satisfies $U_{II} = \lim_{\epsilon\to 0^+} U^\epsilon$, but $U_I \neq \lim_{\epsilon\to 0+} U^\epsilon$. Since a solution of the conservation law can be seen as the derivative of the solution of a Hamilton-Jacobi equation, the same technique of viscosity solutions can be applied to

$$V_t^\epsilon + \frac{(V_x^\epsilon)^2}{2} = \epsilon V_{xx}^\epsilon, \quad \epsilon > 0. \tag{9.18}$$

The functions $V_I(x,t) = -\int_x^\infty U_I(y,t)dy$, and $V_{II}(x,t) = -\int_x^\infty U_{II}(y,t)dy$ have the same initial data and they are both candidates of solutions to the Hamilton-Jacobi equation

$$V_t + \frac{(V_x)^2}{2} = 0.$$

The shock waves for conservation laws corresponds to solutions with discontinuities in the derivative for Hamilton-Jacobi solutions. Only the function V_{II} satisfies

$$V_{II} = \lim_{\epsilon\to 0+} V^\epsilon, \tag{9.19}$$

but $V_I \neq \lim_{\epsilon\to 0+} V^\epsilon$. It can be shown that the condition (9.19) implies uniqueness for Hamilton-Jacobi equations. Note that (9.19) corresponds to the limit of vanishing noise in control of stochastic differential equations.

9.5 Numerical Approximations of Conservation Laws and Hamilton-Jacobi Equations

We have seen that the viscous problem

$$\begin{aligned}&\partial_t u^\varepsilon + \partial_x H(u^\varepsilon) = \varepsilon u^\varepsilon_{xx} \quad \text{for } \ (x,t) \in \mathbb{R} \times (0,+\infty),\\ &u(x,0) = u^0(x) \qquad \text{for } \ x \in \mathbb{R},\end{aligned} \tag{9.20}$$

can be used to construct unique solutions to the conservation law

$$\begin{aligned}&\partial_t u + \partial_x H(u) = 0 \qquad \text{for } \ (x,t) \in \mathbb{R} \times (0,+\infty),\\ &u(x,0) = u^0(x) \qquad \text{for } \ x \in \mathbb{R}.\end{aligned} \tag{9.21}$$

In this section we will develop numerical approximations to the conservation law (9.21) and the related Hamilton-Jacobi equation

$$\partial_t v + H(\partial_x v) = 0,$$

based on viscous approximations. We will also see that too little viscosity may give unstable approximations.

To show the difficulties to solve numerically a problem like (9.21) and (9.20) we consider a related steady-state problem (i.e. a problem that has no dependence on t)

$$\begin{aligned}&\partial_x w(x) - \varepsilon \ \partial_x^2 w(x) = 0 \quad \text{for } x < 0,\\ &\lim_{x\to-\infty} w(x) = 1, \ w(0) = 0,\end{aligned} \tag{9.22}$$

where $\varepsilon \geq 0$ is fixed. It is easy to verify that the exact solution is $w(x) = 1 - \exp(\frac{x}{\varepsilon})$, for $x \leq 0$. Now, we construct a uniform partition of $(-\infty, 0]$ with nodes $x_j = j\Delta x$ for $j = 0, -1, -2, \ldots$, where $\Delta x > 0$ is a given mesh size. Denoting by W_j the approximation of $w(x_j)$, the use of a second order accurate finite element method or finite difference scheme method leads to the scheme

$$\begin{aligned}\frac{W_{j+1} - W_{j-1}}{2\Delta x} - \varepsilon \ \frac{W_{j+1} - 2W_j + W_{j-1}}{(\Delta x)^2} &= 0, \ j = -N+1, \ldots, -1,\\ W_0 &= 0,\\ W_{-N} &= 1.\end{aligned} \tag{9.23}$$

Assume that N is odd. If $\varepsilon \ll \Delta x$, the solution of (9.23) is approximated by

$$\frac{W_{j+1} - W_{j-1}}{2\Delta x} = 0,$$

which yields the oscillatory solution $W_{2i} = 0$ and $W_{2i+1} = 1$ that does not approximate w, instead $\|w - W\|_{L^2} = \mathcal{O}(1)$. One way to overcome this difficulty is to replace, in (9.23), the *physical diffusion* ε by the *artificial diffusion* $\hat{\varepsilon} = \max\{\varepsilon, \frac{\Delta x}{2}\}$. For the general problem $\beta \cdot \nabla u - \varepsilon \Delta u = f$ take $\hat{\varepsilon} = \max\{\varepsilon, |\beta| \frac{\Delta x}{2}\}$. Now, when $\varepsilon \ll \Delta x$, we have $\hat{\varepsilon} = \frac{\Delta x}{2}$ and the method (9.23), with ε replaced by $\hat{\varepsilon}$, yields $W_j = W_{j-1}$ for $j = -N+1, \ldots, -1$, i.e. $W_j = 1$ for $j = -N, \ldots, -1$, which is an acceptable solution with $\|w - W\|_{L^2} = \mathcal{O}(\sqrt{\Delta x})$. Another way to cure the problem is to resolve by choosing Δx enough small, so that $\hat{\varepsilon} = \varepsilon$.

The Lax-Friedrich method for the problem (9.21), is given by

$$U_j^{n+1} = U_j^n - \Delta t \left[\frac{H(U_{j+1}^n) - H(U_{j-1}^n)}{2\Delta x} - \frac{(\Delta x)^2}{2\Delta t} \; D_+ D_- U_j^n \right], \tag{9.24}$$

with

$$D_+ V_j = \frac{V_{j+1} - V_j}{\Delta x}, \; D_- V_j = \frac{V_j - V_{j-1}}{\Delta x} \text{ and } D_+ D_- V_j = \frac{V_{j+1} - 2V_j + V_{j-1}}{(\Delta x)^2}.$$

The stability condition for the method (9.24) is

$$\lambda \equiv \frac{\Delta x}{\Delta t} > \max_u |H'(u)|. \tag{9.25}$$

Note that the method (9.24) is a second order accurate central difference approximation of the equation (9.20) with $\varepsilon = \frac{\lambda}{2} \Delta x$.

We want to approximate the viscosity solution of the one-dimensional Hamilton-Jacobi equation

$$\partial_t v + H(\partial_x v) = 0, \tag{9.26}$$

where $v = \lim_{\varepsilon \to 0+} v^\varepsilon$ and

$$\partial_t v^\varepsilon + H(\partial_x v^\varepsilon) = \varepsilon \;\; \partial_x^2 v^\varepsilon. \tag{9.27}$$

Setting $u = \partial_x v$ and taking derivatives in (9.26), we obtain a conservation law for u, i.e.

$$\partial_t u + \partial_x H(u) = 0. \tag{9.28}$$

To solve (9.26) numerically, a basic idea is to apply (9.24) on (9.28) with $u_i^n = (V_{i+1}^n - V_{i-1}^n)/(2\Delta x)$ and then use summation over i to approximate the integration in (9.11). We get

$$\frac{V_{j+1}^{n+1} - V_{j-1}^{n+1}}{2\Delta x} = \frac{V_{j+1}^n - V_{j-1}^n}{2\Delta x}$$
$$-\Delta t\Bigg[\frac{H\Big(\frac{V_{j+2}^n - V_j^n}{2\Delta x}\Big) - H\Big(\frac{V_j^n - V_{j-2}^n}{2\Delta x}\Big)}{2\Delta x} - \frac{(\Delta x)^2}{2\Delta t} \; D_+D_-\frac{V_{j+1}^n - V_{j-1}^n}{2\Delta x}\Bigg].$$

Summing over j and using that $V_{-\infty}^m = 0$ and $H(0) = 0$, it follows that

$$V_j^{n+1} = V_j^n - \Delta t\Bigg[H\Big(\frac{V_{j+1}^n - V_{j-1}^n}{2\Delta x}\Big) - \frac{(\Delta x)^2}{2\Delta t}D_+D_-V_j^n\Bigg], \tag{9.29}$$

which is the Lax-Friedrich method for (9.26). Note that (9.29) is a second order accurate central difference approximation of the equation

$$\partial_t v + H(\partial_x v) = \frac{(\Delta x)^2}{2\Delta t}\big(1 - (\frac{\Delta t}{\Delta x}H')^2\big)\partial_x^2 v,$$

which is (9.27) with artificial diffusion $\frac{\lambda\Delta x}{2}$.

In the two-dimensional case a first order Hamilton-Jacobi equation has the form

$$\partial_t v + H(\partial_{x_1} v, \partial_{x_2} v) = 0. \tag{9.30}$$

The analogous scheme to (9.29) for that equation is

$$\begin{aligned} V_{j,k}^{n+1} = V_{j,k}^n - \Delta t\Bigg[\quad & H\Big(\frac{V_{j+1,k}^n - V_{j-1,k}^n}{2\Delta x_1}, \frac{V_{j,k+1}^n - V_{j,k-1}^n}{2\Delta x_2}\Big) \\ & -\frac{(\Delta x_1)^2}{4\Delta t}\frac{V_{j+1,k}^n - 2V_{j,k}^n + V_{j-1,k}^n}{(\Delta x_1)^2} \\ & -\frac{(\Delta x_2)^2}{4\Delta t}\frac{V_{j,k+1}^n - 2V_{j,k}^n + V_{j,k-1}^n}{(\Delta x_2)^2} \quad \Bigg] \end{aligned}$$

which for $\Delta x_1 = \Delta x_2 = h$ and $\lambda = h/\Delta t$ corresponds to a second order approximation of the equation

$$\partial_t v^h + H(\partial_{x_1} v^h, \partial_{x_2} v^h) = \frac{\Delta x^2}{4\Delta t}\sum_i \partial_{x_i x_i} v - \sum_{i,j}\frac{\Delta t}{2}\partial_{x_i} H \partial_{x_j} H \partial_{x_i x_j} v.$$

9.6 Symmetric Hyperbolic Systems

We consider the following system of partial differential equations: find $u : \mathbb{R}^d \times [0, +\infty) \to \mathbb{R}^n$ satisfying

$$Lu \equiv A_0 \partial_t u + \sum_{i=1}^{d} A_i \, \partial_{x_i} u + B \, u = f \text{ on } \mathbb{R}^d \times (0, +\infty), \quad (9.31)$$
$$u(x, 0) = u_0(x) \text{ for } x \in \mathbb{R}^d,$$

where $u_0 : \mathbb{R}^d \to \mathbb{R}^n$ is given initial data, $\{A_i\}_{i=0}^d$ and B are given $n \times n$ matrices, and f is a given n vector. We say that the system (9.31) is a *symmetric hyperbolic system*, when the matrices $\{A_i\}_{i=0}^d$ are symmetric and the matrix A_0 is positive definite. We note that, in general, $A_i = A_i(x, t, u)$, $i = 0, \ldots, d$, $B = B(x, t, u)$ and $f = f(x, t, u)$, and only in the linear case we have $A_i = A_i(x, t)$, $i = 0, \ldots, d$, $B = B(x, t)$ and $f = f(x, t)$. On what follows denote the Euclidean inner product $x \cdot y = \sum_{i=1}^d x_i y_i$. Besides this, we assume the linear case, and in addition that

- A_0 is uniformly positive definite, i.e. there are positive constants μ_a, μ_b such that

$$\mu_a |x|^2 \leq \; A_0 x \cdot x \; \leq \mu_b |x|^2 \quad \forall \, x \in \mathbb{R}^d, \; \forall t \in [0, T], \quad (9.32)$$

- all the matrices $\partial_t A_0$, B, A_i, $\partial_i A_i$, $i = 1, \ldots, d$ are uniformly bounded, i.e. there exist a positive constant K such that if A is any of these matrices then

$$|Ax \cdot x| \leq K |x|^2 \quad \forall \, x \in \mathbb{R}^d, \; \forall t \in [0, T]. \quad (9.33)$$

Example 9.12 We consider the initial value problem for a multidimensional wave equation

$$\partial_t^2 u = \sum_{i,k=1}^{d} a_{ik} \partial_{x_i x_k} u + \sum_{i=1}^{d} b_i \partial_{x_i} u + c_0 \partial_t u + d_0 \;\; u, \quad (9.34)$$
$$u(x, 0) = u_0(x) \text{ and } \partial_t u(x, 0) = u_1(x) \text{ for } x \in \mathbb{R}^d,$$

where the matrix $(a_{ik})_{i,k=1}^{d}$ is symmetric and positive definite. Introducing the new variables

$$v_i = \partial_{x_i} u,\ i = 1, \ldots, d,\ v_{d+1} = \partial_t u,\ v_{d+2} = u,$$

the equation (9.34) is written equivalently as

$$\begin{aligned}
&\sum_{k=1}^{d} a_{ik}\ \ \partial_t v_k - \sum_{k=1}^{d} a_{ik}\ \ \partial_{x_k} v_{d+1} = 0,\ i = 1, \ldots, d,\\
&\partial_t v_{d+1} - \sum_{i,k=1}^{d} a_{ik}\ \ \partial_{x_k} v_i - \sum_{i=1}^{d} b_i\ \ v_i - c_0\ \ v_{d+1} - d_0\ \ v_{d+2} = 0,\\
&\partial_t v_{d+2} - v_{d+1} = 0,\\
&v(x, 0) = (\partial_{x_1} u_0(x), \ldots, \partial_{x_d} u_0(x), u_0(x), u_1(x)) \text{ for } x \in \mathbb{R}^d.
\end{aligned}$$

Defining the matrices

$$A_0 = \begin{bmatrix} a_{11} & \ldots & a_{1d} & 0 & 0 \\ \vdots & & \vdots & \vdots & \vdots \\ a_{d1} & \ldots & a_{dd} & 0 & 0 \\ 0 & \ldots & 0 & 1 & 0 \\ 0 & \ldots & 0 & 0 & 1 \end{bmatrix}, \quad A_k = \begin{bmatrix} 0 & \ldots & 0 & -a_{1k} & 0 \\ \vdots & & \vdots & \vdots & \vdots \\ 0 & \ldots & 0 & -a_{dk} & 0 \\ -a_{1k} & \ldots & -a_{dk} & 0 & 0 \\ 0 & \ldots & 0 & 0 & 0 \end{bmatrix}, \quad k = 1, \ldots, d,$$

and

$$B = \begin{bmatrix} 0 & \ldots & 0 & 0 & 0 \\ \vdots & & \vdots & \vdots & \vdots \\ 0 & \ldots & 0 & 0 & 0 \\ -b_1 & \ldots & -b_d & -c_0 & -d_0 \\ 0 & \ldots & 0 & -1 & 0 \end{bmatrix},$$

Equation (9.35) takes the form

$$A_0 \partial_t v + \sum_{k=1}^{d} A_k \partial_{x_k} v + Bv = 0$$

which is a symmetric hyperbolic system, since $\{A_i\}_{i=0}^{d}$ are symmetric and A_0 is positive definite. $\square$

Exercise 9.13 Show that the Maxwell Equations in the vacuum

$$\varepsilon \frac{\partial E}{\partial t} - \mathrm{rot} H = 0,$$
$$\mu \frac{\partial H}{\partial t} + \mathrm{rot} E = 0,$$

where $E(x,t)$ and $H(x,t) \in \mathbb{R}^3$, can be written as a symmetric hyperbolic system.

Example 9.14 Consider the symmetric hyperbolic system in one space dimension

$$\partial_t u + A\partial_x u + Bu = f, \quad t > 0, \ x \in \mathbb{R}$$
$$u(x,0) = u_0(x) \text{ for } x \in \mathbb{R},$$

where A is a $n \times n$ symmetric matrix. Since A is symmetric it can be diagonalized by via an orthogonal matrix P ($P^{-1} = P^T$), i.e.

$$P^{-1}AP = \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n).$$

Let $w = P^{-1}u$. Then, it follows that

$$\begin{aligned} f = & \quad \partial_t u + A\partial_x u + Bu \\ = & \quad Pw_t + AP\partial_x w + (P_t + AP_x + BP)w \end{aligned}$$

which yields

$$\begin{aligned} & w_t + \Lambda\partial_x w + \Pi w = g \\ & w(x,0) = w_0(x) \equiv P^{-1}u_0(x,0) \text{ for } x \in \mathbb{R}, \end{aligned} \tag{9.35}$$

where

$$\Pi \equiv P^{-1}(P_t + AP_x + BP) \text{ and } g = P^{-1}f.$$

When the matrix Π is diagonal, i.e., $\Pi = \mathrm{diag}(\pi_1, \ldots, \pi_n)$, we can solve equation (9.35) by the method of characteristics $\{X_j\}_{j=1}^n$. A characteristic is the solution of the following the system of ordinary differential equations

$$\begin{aligned} & \frac{d}{dt}X_j(x_*;t) = \lambda_j(X_j(x_*;t),t) \quad \text{for } t > 0, \ j = 1, \ldots, n, \\ & X_j(x_*;0) = x_*, \end{aligned}$$

for a given $x_* \in \mathbb{R}$. Solving the equations for the characteristics above we can obtain a solution of the initial problem (9.35), since we have for $j = 1, \dots, n$

$$\begin{aligned} \frac{d}{ds} w_j(X_j(x;s),s) &= \partial_t w_j(X_j(x;s),s) + \lambda_j(X_j(x;s),s) \;\; \partial_{x_j} w(X_j(x;s),s) \\ &= -\pi_j(X_j(s),s) \;\; w_j(X_j(x;s),s) + g_j(X_j(x;s)), \quad (9.36) \end{aligned}$$

for $s > 0, \ x \in \mathbb{R}$.

Exercise 9.15 Generalize (9.36) to the symmetric hyperbolic system $A_0 \partial_t u + A_1 \partial_x u + Bu = f$.

Example 9.16 Let us apply the procedure above to the special case of the wave equation in one space dimension

$$\begin{aligned} &\partial_t^2 v - \partial_x^2 v = 0, \\ &v(x,0) = v_0(x), \;\; \partial_t v(x,0) = v_1(x) \quad \text{for } \; x \in \mathbb{R}. \end{aligned}$$

Setting $u_1 = \partial_t v$, $u_2 = \partial_x v$ and $u = (u_1, u_2)$, the equation above is

$$\begin{aligned} &\partial_t u + M \partial_x u = 0 \\ &u(x,0) = (v_1(x), \partial_x v_0(x)) \text{ for } x \in \mathbb{R}, \end{aligned}$$

where

$$M = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}.$$

The matrix M has eigenvalues ± 1 with corresponding eigenvectors $(1, -1)$ and $(1, 1)$. The matrix P that diagonalizes M is

$$P = \frac{\sqrt{2}}{2} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}, \text{ i.e. } P^T M P = \operatorname{diag}(1, -1).$$

Then we set $w = P^{-1} u$ and obtain $w_1(x+t, t) = w_1(x, 0)$ and $w_2(x-t, t) = w_2(x, 0)$. □

Next we will prove an *energy estimate* which provides a uniqueness result for a linear symmetric hyperbolic system. These energy estimate is also the

Figure 9.5: Space-Time domain Ω where Gauss theorem is applied to obtain an energy estimate.

basis for an existence proof, first given by Friedrichs based on the construction in Lax-Milgram's Theorem. We multiply the linear system (9.31) by u to get

$$\partial_t(A_0 u \cdot u) + \sum_{i=1}^{d} \partial_{x_i}(A_i u \cdot u) + Cu \cdot u = 2f \cdot u \tag{9.37}$$

where $x \cdot y = \sum_{i=1}^{d} x_i y_i$ is the standard inner product in $\mathbb{R}^d$ and

$$C = 2B - \partial_t A_0 - \sum_{i=1}^{d} \partial_{x_i} A_i. \tag{9.38}$$

Let us consider a "space-time" cone $\Omega \subset [0,T] \times \mathbb{R}^d$ as shown in Figure 9.5. Observe that $\partial\Omega = \Omega_T \cup \Omega_L \cup \Omega_0$ and use the notation $\Omega_t \equiv \{(x,s) \in \Omega : s = t\}$ to describe "time slices" of the domain Ω. Integrating (9.37) in Ω and using Gauss theorem in Ω yields

$$\begin{aligned} \int_{\Omega_T} (A_0 u, u)dx - \int_{\Omega_0} (A_0 u, u)dx + \int_{\Omega_L} (\sum_{i=1}^{d} \boldsymbol{n}_i A_i u \cdot u)dS \\ = \int_{\Omega} \big[2f \cdot u - Cu \cdot u\big] dxdt. \end{aligned} \tag{9.39}$$

Recall that the matrix A_0 is positive definite, so there are positive constants μ_a, μ_b and k_0, such that

$$\mu_a |x|^2 \leq A_0 x \cdot x \leq \mu_b |x|^2 \quad \forall\, x \in \mathbb{R}^d.$$

Besides this, A_i, $i = 1, \dots, d$ are both symmetric and uniformly bounded, so the matrix

$$\boldsymbol{n}_0 A_0 + \sum_{i=1}^{d} \boldsymbol{n}_i A_i$$

is symmetric positive definite as long as the lateral boundary of the cone, Ω_L, is chosen such that its unitary normal vector satisfies

$$n_0 > \frac{K\sqrt{d}}{\mu_a} \sqrt{\sum_{i=1}^{d} n_i^2}.$$

Thus, with this choice of the cone Ω, (9.39) yields

$$E(T) \leq E(0) + \int_0^T \left(\int_{\Omega_t} \big[2f \cdot u - Cu \cdot u \big] dx \right) dt, \tag{9.40}$$

where E is the energy of the system (9.31) defined by

$$E(t) \equiv \int_{\Omega_t} A_0(x,t) u(x,t) \cdot u(x,t) dx \quad \text{for } t \geq 0.$$

Now we estimate the right hand side of (9.40) using the coercitivity of A_0 (9.32) and the boundedness (9.33),(9.38), of C, to find a constant k_0 such that

$$|Cu \cdot u| \leq k_0 A_0 u \cdot u,$$

and

$$\begin{aligned} 2|f \cdot u| \leq & \quad \mu_a |u|^2 + \frac{1}{\mu_a} |f|^2 \\ \leq & \quad A_0 u \cdot u + \frac{1}{\mu_a} |f|^2. \end{aligned}$$

From the relations above, we arrive at

$$E(t) \leq E(0) + (1 + k_0) \int_0^t E(s) ds + \frac{1}{\mu_a} \int_0^t \int_{\Omega_s} |f(x,s)|^2 dx ds, \ \forall t \in [0,T],$$

and consequently the Grönwall's lemma 3.2 yields

$$\sup_{s \in [0,T]} E(s) \leq \exp((k_0+1)T) \ \left[E(0) + \frac{1}{\mu_a} \int_0^T \int_{\Omega_s} |f(x,s)|^2 dx ds \right], \tag{9.41}$$

which is the basic stability estimate for the linear symmetric hyperbolic system (9.31).

Remark 9.17 [Cones of influence and dependence] Observe that the energy estimate implies that perturbations travel with finite speed if they satisfy a linear symmetric hyperbolic system.

Remark 9.18 [Uniqueness] Using the stability estimate (9.41), we can prove uniqueness of the solution to symmetric hyperbolic systems: let u_1 and u_2 be solutions of a symmetric hyperbolic system with the same initial condition. Then the function $u = u_1 - u_2$ solves the linear initial value problem

$$\begin{aligned} Lu &= 0 \text{ on } \mathbb{R}^d \times (0,T), \\ u(x,0) &= 0 \text{ for } x \in \mathbb{R}^d. \end{aligned}$$

Applying the stability estimate (9.41) in Ω we obtain that the energy of each slice Ω_t satisfies $E(t) = 0$ for $t \in [0,T]$, so $u = 0$ in Ω and hence $u_1 = u_2$ in Ω. The fact that any bounded region in $\mathbb{R}^d \times [0,T]$ is included in a cone like Ω finishes the argument.

Remark 9.19 [Estimate] For the particular case with zero initial datum, i.e. $u(\cdot,0) = 0$, the energy estimate (9.41) implies

$$\|u\|_{L^2(\Omega)} \le \Gamma_T \|f\|_{L^2(\Omega)} = \Gamma_T \|Lu\|_{L^2(\Omega)}. \tag{9.42}$$

with $\Gamma_T^2 \equiv T\exp((1+k_0)T)/\mu_a^2$.

Theorem 9.20 (Existence result for Hyperbolic equations) *There exists a solution of a symmetric hyperbolic system that satisfies the assumptions (9.32 and 9.33).*

Proof. The proof is divided in three steps, namely

1. define the notion of weak solutions for the symmetric hyperbolic system,

2. define the adjoint operator and the dual problem, and then

3. apply the Lax Milgram's theorem to an auxiliary equation which is related to the original symmetric hyperbolic system.

Step 1: Without loss of generality, let us assume zero initial data, i.e. $u_0 = 0$. In order to set the notion of weak solutions, we need to introduce the concept of test functions, which in this case is given by the set

$$\tilde{V} = \begin{array}{l} \{v : \mathbb{R}^d \times [0,T] \to \mathbb{R}^d \ : \ v \in C^1([0,T] \times \mathbb{R}^d), \\ \qquad\qquad\qquad\qquad v \text{ with compact support and } v(T,\cdot) = 0\}. \end{array}$$

Step 2: Let us define the adjoint operator by the identity

$$(Lu, v) = \int_0^T \int_{\mathbb{R}^d} (Lu) \cdot v dx\ dt = (u, L^* v), \forall v \in \tilde{V}, \tag{9.43}$$

with the inner product $(v,w) = \int_0^T \int_{\mathbb{R}^d} v \cdot w dx\ dt$. Using the boundary conditions $u(0,\cdot) = v(T,\cdot) = 0$ and integrating by parts the last equation yields

$$L^* v = -\partial_t (A_0 v) - \sum_{i=1}^d \partial_i (A_i v) - B^T v = -Lv + (C^T - B^T + 2B)v.$$

Observe then that $-L^*$ is a symmetric hyperbolic operator, since the operator L is symmetric hyperbolic. With the help of the adjoint operator we say that $u \in L^2(\mathbb{R}^d \times [0,T])$ is a weak solution of

$$\begin{cases} \quad Lu = f \text{ on } \mathbb{R}^d \times (0,T), \\ u(0,\cdot) = 0 \text{ on } \mathbb{R}^d, \end{cases} \tag{9.44}$$

if

$$(u, L^* v) = (f, v), \ \forall v \in \tilde{V}, \tag{9.45}$$

something that clearly holds for a classical solution. We leave as an exercise to verify that (9.45) and the assumption that u is a smooth function imply that u is a classical smooth solution of (9.44).

Step 3: To conclude we use the Lax Milgram's theorem. Let us consider the bilinear form

$$B(u,v) \equiv (L^* u, L^* v)$$

and a Hilbert space H which has the bilinear form B as its inner product and norm $\|v\|_H \equiv B(v,v)^{1/2}$, so that H is the completion of $\tilde{V}$ with respect to the norm $\|\cdot\|_H$. Let us consider the following auxiliary equation

$$B(z,v) = \mathcal{L}(v), \ \forall v \in H,$$

with $\mathcal{L}(v) \equiv (f,v)$ and prove that there exist a solution $z \in H$ for it. To this end, the linear functional $\mathcal{L} : H \to \mathbb{R}$ must be bounded in H, i.e. there exist a real constant $0 < C < \infty$ such that

$$|\mathcal{L}(v)| \leq C\|v\|_H, \ \forall v \in H.$$

To verify this estimate use the fact that the operator $-L^*$ is also symmetric hyperbolic as L, so an application of the estimate (9.42) gives

$$\|v\|_{L^2(\Omega)} \leq \Gamma'_T\|L^*v\|_{L^2(\Omega)} \leq \Gamma'_T\|v\|_H, \ \forall v \in H.$$

Furthermore, since the cone Ω can be arbitrary large and the space H is the completion of functions with compact support in $\mathbb{R}^d \times [0,T]$ we have

$$\|v\|_{L^2(\mathbb{R}^d\times[0,T])} \leq \Gamma'_T\|v\|_H, \ \forall v \in H.$$

Then the functional $\mathcal{L}$ is bounded in H with $C = \Gamma'_T\|f\|_{L^2(\mathbb{R}^d\times[0,T])}$ because

$$|\mathcal{L}(v)| \leq \|f\|_{L^2(\mathbb{R}^d\times[0,T])}\|v\|_{L^2(\mathbb{R}^d\times[0,T])} \leq \|f\|_{L^2(\mathbb{R}^d\times[0,T])}\Gamma'_T\|v\|_H, \ \forall v \in H.$$

Thus, from the Lax Milgram's theorem applied to the space H with norm $\|\cdot\|_H$ there exist $z \in H$ such that

$$B(z,v) = \mathcal{L}(v), \ \forall \ v \in H$$

and consequently $u \equiv L^*z \in L^2(\mathbb{R}^d \times [0,T])$ is, according to (9.45) a weak solution of the symmetric hyperbolic equation since

$$(v,f) = \mathcal{L}(v) = B(z,v) = (L^*z, L^*v) = (u, L^*v), \forall v \in H.$$

□

Exercise 9.21 [ ] Under the hypothesis of the previous theorem, show that

$$\|u\|_{L^2(\mathbb{R}^d\times[0,T])} \leq \Gamma'_T\|f\|_{L^2(\mathbb{R}^d\times[0,T])}.$$

□

Exercise 9.22 [ ] Show that $B(\cdot,\cdot)$ is a bilinear symmetric and L^2-elliptic functional which is not L^2 continuous.

□

Chapter 10

References

The following references have been useful for preparing these notes and are recommended for further studies.

10.1 Stochastic Differential Equations

[KP]*Numerics for SDE:* Peter E. Kloeden and Eckhard Platen, Numerical Solution of Stochastic Differential Equations, *Springer*, (1992).

[Mi]*Numerics for SDE:* G. N. Milstein, Numerical Integration of Stochastic Differential Equations, Mathematics and its applications, vol. 313, *Kluwer Academic Publishers*, 1994.

[O] *SDE:* Bernt Øksendal, Stochastic Differential Equations - An Introduction with Applications, *Springer*, (1998).

[KS] *Advanced SDE:* Ioannis Karatzas and Steven E. Shreve, Brownian Motion and Stochastic Calculus, *Springer*, (1999).

10.2 Probability

[D] Richard Durrett, Probability: theory and examples. Second edition. Duxbury Press, Belmont, CA, 1996.

10.3 Mathematical Finance

[BR] *Basic stochastics for finance:* M. Baxter and A. Rennie, Financial Calculus: An introduction to derivate pricing, *Cambridge University Press*, (1996).

[H] *Finance in practice:* John C. Hull, Options, futures and other derivatives, *Prentice Hall*, (1997).

[WHD] *Finance with numerics:* Paul Wilmott, Sam Howison and Jeff Dewynne, The Mathematics of Financial Derivatives, *Cambridge University Press*, (1995).

10.4 Partial Differential Equations

[E] *Advanced PDE:* Lawrence C. Evans, Partial Differential Equations, *AMS*, GSM 19, (1998).

[J] *FEM:* Claes Johnson, Numerical solution of partial differential equations by the finite element method, *Studentlitteratur*, (1987).

[BS] *Advanced FEM:* Susanne C. Brenner and L. Ridgway Scott, The Mathematical Theory of Finite Element Methods, *Springer*, TAM 15, (1994).

[EEHJ] *Introductory DE and PDE:* K. Eriksson, D. Estep, P. Hansbo and C. Johnson, Computational Differential Equations, *Cambridge University Press*, (1996).

[S] *Introductory DE and PDE:* Gilbert Strang, Introduction to Applied Mathematics, *Wellesley-Cambridge Press*, (1986).

10.5 Variance Reduction for Monte Carlo Methods

[C] Russel E. Calflisch, Monte Carlo and quasi-Monte Carlo methods, *Acta Numerica*, 1-49, (1998).