MATHEMATICS
of
COMPUTATIONAL FINANCE

Lecture Notes

Andrzej Palczewski

University of Warsaw 2022

Contents

Preface

The goal of these lecture notes is to give a relatively short but sufficiently rigorous, and hopefully readable account of basic numerical methods used in pricing derivative financial products. Derivative pricing is one of the many problems of financial mathematics. Rigorous pricing methods started in the seventies with the celebrated model of Black, Scholes, and Merton and became a basis for the pricing of thousands of complex financial products. There is a large number of excellent books dealing with numerical implementations of these pricing methods. But the majority of them are focused on numerical algorithms used in pricing. These notes have originated from a course given to students of mathematics at the University of Warsaw. That audience obliged the lecturer to speak not only about algorithms but also about the mathematical foundations which stay behind these algorithms: convergence, stability, and error estimates. The main body of the text is restricted to theoretical material with a limited number of examples related to financial models.

Writing these notes I have assumed that the reader possesses a broad knowledge of continuous finance and its mathematical models. Thus there are no introductory chapters on quantitative finance. All described numerical methods start with the presentation of the analytical problem as a stochastic or partial differential equation without discussing the relation to a financial model. But the numerical methods are quite general and the experienced reader can recognize that they cover not only the Black-Scholes model but also local and stochastic volatility models. On the other hand, the reader whose acquaintance with quantitative finance is limited to the Black-Scholes model can ignore the multidimensional approach since the results are also valid for simple models.

The fundamental principle of numerical analysis is that one can compute only solutions that do exist. Following that principle, I have included in these notes theorems that guarantee the existence (and possibly uniqueness) of solutions to equations that are subjects of forthcoming numerical analysis. That goal is easy to achieve for financial models formulated in probabilistic terms (random variables, stochastic processes, stochastic differential equations) because they are standard tools used in continuous finance. Besides, the most advanced result I am using, is

the existence and properties of solutions for the stochastic differential equation of Itô type, the results belonging to the usual prerequisites for any course of quantitative finance. The situation is completely different with financial models described by partial differential equations (PDE). First, a course in the PDE theory is rarely required as a prerequisite for financial courses. Second, to formulate a theorem on the existence of solutions for a PDE model, we need pretty advanced tools: weak derivatives, Sobolev's spaces, weak solutions, etc. Therefore, I have decided to thoroughly introduce partial differential equations together with the Feynman-Kac theorem, which gives a rigorous passage from the stochastic to the PDE formulation of financial models. Finally, presenting existence proofs for PDEs has some didactic aspects, as these proofs are in many cases prototypes for proofs of convergence for numerical methods. Hence, before passing to operations with multiple indices of a numerical algorithm, the reader can see the proof idea in a clean functional space formulation.

Since the notes are written mainly for students of mathematics or mathematically oriented students of economics or natural sciences who study quantitative finance for academic or professional purposes the presentation is quite advanced taking for granted a broad knowledge of analysis, probability, and statistics with some orientation in stochastic processes. The whole presentation is restricted to pricing derivative instruments in a financial model described by the Itô SDE and the corresponding parabolic PDE. This is a simplification that makes it easier to achieve the main goal of presenting complete proofs. These complete proofs are compilations from many sources, sometimes with my original additions. There are no references to these sources in the body of the text. But the list of references contains books and research papers from which I have profited writing these notes. Of course, I was not able to succeed with writing proofs of all theorems. There are many theorems without proof. For such theorems, I usually quote the reference in which the proof can be found. Essentially, I omit proofs that do not belong to the area of numerical analysis and the reason for skipping them is twofold: first, some proofs belong to the field of mathematics very distant from the topic of these notes, and presenting them will require plenty of auxiliary material (this is particularly visible in the chapter on American options); second, some proofs need advanced tools being far beyond the knowledge which I can expect from the reader. In some places, I have made a compromise by just quoting a result necessary in the proof without mentioning even where such a result comes from. Presenting the theorem with proofs I formulate the theorem possibly in full generality and then write the proof of the most elementary case: one-dimensional, with constant coefficients, smooth data, etc. The reason for such an approach is purely didactic: I intend to present to the reader the main idea which stays behind the proof and this main idea can be lost in the orgy of multiple indices, splittings into subdomains, smoothing

data and coefficients, etc., required for the proof of the sharp version.

The interest of the notes lies in the computation of derivative prices. Since financial models rely on stochastic differential equations, applying Monte-Carlo methods to derivative pricing is very natural. Similarly, tree methods are very intuitive and fast. Applying the Feynman-Kac theorem, we can reduce derivative pricing to solutions of partial differential equations which gives rise to a large class of numerical methods for PDEs. All these computational approaches are described in these lecture notes. But the presentation is far from being complete. After all, there are just lecture notes and not a monograph and the author has selected for presentation the most popular methods and algorithms. Of course, the selection is biased by the author's experience in computational finance. The algorithms and methods of these notes have been practically tested by myself or my students during many years of lecturing computational finance. The experienced reader can ask why the notes are limited to models in which Wiener processes describe randomness and there are no models with jumps and Lévy processes. The reasons are numerous. One of these is that including Lévy processes will require writing a thorough monograph instead of medium-size lecture notes. Such a monograph is beyond the aspiration of the author.

The notes are organized as follows. After the introductory Chapter 1, in which binomial and trinomial trees are presented briefly, Chapter 2 addresses the generation of pseudo-random numbers. It is explained in this chapter how samples from a given distribution can be generated using pseudo-random numbers. Algorithms for the generation of normal deviates get particular attention in that chapter. Chapter 3 starts with crude Monte Carlo. Then, the variance reduction technique is presented. The chapter concludes with the computation of Greeks. In Chapter 4, numerical solutions of stochastic differential equations are discussed with proofs of convergence for the Euler and Milstein schemes and error estimates.

A large part of the notes is devoted to numerical solutions of partial differential equations arising in finance. As an introduction to the topic, Chapter 5 collects fundamental facts from the theory of weak solutions of PDEs in Sobolev's spaces. Chapter 6 focuses on finite difference methods for the partial differential equations of parabolic type. The most popular finite difference schemes are described and their accuracy (order of approximation) and stability are proved. Finite element methods are treated in Chapter 7 for both elliptic and parabolic problems. It is proved in this chapter that finite element approximations converge to the corresponding solutions of differential problems. Chapter 8 is devoted to American options. Both Monte Carlo and PDEs methods of pricing are presented. The Monte Carlo approach is limited to the description of the frequently used algorithm of Longstaff and Schwartz. Then a careful analysis of its convergence is provided. In the final part of the chapter, the variational approach to pricing American options is

discussed. After a short introduction to variational inequalities related to American options, the presentation is concentrated on two popular numerical methods: the projected SOR and the penalty method. For both of these methods, the proofs of stability and convergence are given. As I have mentioned before, the notes originated from a course in computational finance. But definitely, the scope of these notes is too broad to fit into a single course. A reasonable selection of material from Chapters 2–6 can be lectured in one course. On the other hand, the chapter on American options supplemented by more information on optimal stopping problems and variational inequalities in continuous time can be sufficient for a special course.

I would like to express my gratitude to several people who have influenced me on these lecture notes. I owe a particular debt to Piotr Kowalczyk with whom I have been lecturing computational finance to several cohorts of students at the University of Warsaw. I should also like to acknowledge the students who, during the courses given in the past years, have helped improve the quality of the text through their questions, comments, and feedback.

It is greatly appreciated if the readers could forward any errors, misprints or suggested corrections to A.Palczewski@mimuw.edu.pl

Chapter 1

Introduction

1.1 Financial market

The goal of these lecture notes is the analysis of computational problems of financial derivatives. We start with a formal definition of the financial market in continuous time and finite time horizon T. Uncertainty in the financial market is modeled by a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with a filtration $\mathbb{F} = (\mathcal{F}_t)_{0 \leq t \leq T}$ satisfying the usual conditions of completeness and right-continuity. In addition, we assume that the σ-field $\mathcal{F}_0$ is trivial, i.e. for every $A \in \mathcal{F}_0$ either $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$, and that $\mathcal{F}_T = \mathcal{F}$. The financial market contains $d + 1$ basic traded assets, called underlying securities, whose prices are given by stochastic processes $X_t^0, X_t^1, \ldots, X_t^d$ in $(\Omega, \mathcal{F}, \mathbb{P})$. We assume that these processes are adapted to filtration $\mathbb{F}$, right-continuous with left limits, positive semimartingales.

To account for the time value of money, we introduce a discount process – a *numéraire*.

DEFINITION. 1.1 *A* numéraire *is a stochastic process X_t almost surely strictly positive for each $t \in [0, T]$.*

In what follows, we assume that X_t^0 is a non-dividend paying asset which is almost surely strictly positive for each $t \in [0, T]$, and which we use as numéraire. We will also use the *discount process* $\beta_t = (X_t^0)^{-1}$.

Remark. 1.1 *Usually the money banking account $B_t = e^{\int_0^t r(s)ds}$ with deterministic $r(t)$ (say $r(t) = r$, a constant interest rate) plays the role of numéraire.*

If $B(t, T)$ is the value at $t \leqslant T$ of an asset paying one currency unit at time T, then $B(t, T) = e^{-\int_t^T r(s)ds} = B_t B_T^{-1}$.

Hence, the financial market consists of d price processes $X_t = (X_t^1, \ldots, X_t^d)$ which we call risky assets and the numéraire X_t^0, all defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The dynamics of X_t is given by the system of stochastic differential equations

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dW_t, \quad X_0 = x, \tag{1.1}$$

where b is a d-dimensional vector, σ a matrix of dimension $d \times m$, and W_t is an m-dimensional Wiener process.
Written in the components of vector X equation (1.1) reads

$$dX_t^i = b_i(t, X_t)dt + \sum_{j=1}^{m} \sigma_i^j(t, X_t)dW_t^j, \quad X_0^i = x_i,\ i = 1, \ldots, d.$$

We assume that the coefficients of equation (1.1) fulfill conditions that guarantee the existence of strong solutions.

THEOREM. 1.2 *Let the coefficients in equation* (1.1) *be such that:*

(A1) $|b(t,x) - b(t,y)| + |\sigma(t,x) - \sigma(t,y)| \leq K|x-y|$;

(A2) $|b(t,x)| + |\sigma(t,x)| \leq K(1+|x|)$;

for $x, y \in \mathbb{R}^d$ *and* $t \in [0,T]$, *where* $|\cdot|$ *denotes the Euclidean norm in* $\mathbb{R}^d$ *and* K *is a finite, positive constant.*

If x *is a square integrable random variable (*$\mathbb{E}\big(|x|^2\big) < \infty$*) independent of* W_t *then*

1. *The sequence of iterates*

$$X_t^{(0)} = x,\ X_t^{(n+1)} = x + \int_0^t b\big(s, X_s^{(n)}\big)ds + \int_0^t \sigma\big(s, X_s^{(n)}\big)dW_s$$

converges to X_t.

2. X_t *is a unique, strong solution of the stochastic differential equation*

$$X_t = x + \int_0^t b\big(s, X_s\big)ds + \int_0^t \sigma\big(s, X_s\big)dW_s.$$

3. *The process* X_t *is square-integrable, and for each* $T > 0$ *there exists a constant* C *such that*

$$\mathbb{E}\big(|X_t|^2\big) \leq C\Big(1 + \mathbb{E}\big(|x|^2\big)\Big)\mathrm{e}^{Ct},\ 0 \leq t \leq T.$$

4. *If* $\mathbb{E}(|x|^k) < \infty$, *for some* $k \geq 2$, *then*

$$\mathbb{E}\Big(\sup_{0\leq s\leq t}|X_s|^k\Big) \leq C\Big(1+\mathbb{E}\big(|x|^k\big)\Big),\ 0 \leq t \leq T.$$

5. *If* $\mathbb{E}(|x|^k) < \infty$, *for some* $k \geq 2$, *then*

$$\mathbb{E}\big(|X_t - X_s|^k\big) \leq C\Big(1+\mathbb{E}\big(|x|^k\big)\Big)|t-s|^{k/2},\ |t-s| \leq 1, t, s \in [0,T].$$

Our principal goal is the valuation of contingent claims which are $\mathcal{F}_T$-measurable random variables of the form $g(X_T)$, where X_T is a solution of equation (1.1) taken at time T. Such instruments are called European contingent claims. On a limited scale, we will analyze computational methods for American instruments and, in several examples, instruments with prices depending on the whole trajectory of a process $(X_t)_{0\leq t\leq T}$ (exotic instruments).

We will discuss contingent claim pricing in the idealized market model assuming that the market fulfills the conditions:

1. The market is frictionless: there are no transaction costs, no taxes, costs of borrowing and lending are equal, there are no liquidity restrictions – all assets are accessible in unlimited quantity.

2. Market participants are price takers and are rational: they prefer more to less.

3. The market is arbitrage-free.

To valuate the contingent claim $g(X_T)$ we introduce the notion of a *trading strategy*.

DEFINITION. 1.3 *An* $\mathbb{R}^{d+1}$*- valued predictable, left-continuous process*

$$\varphi_t = (\varphi_t^0, \varphi_t^1, \ldots, \varphi_t^d), \quad t \in [0,T]$$

such that

$$\int_0^T \mathbb{E}(\varphi_t^0)dt < \infty, \quad \sum_{i=1}^d \int_0^T \mathbb{E}\big(|\varphi_t^i|^2\big)dt < \infty$$

is called a trading strategy. Here φ_t^i *is understood as the number of shares of asset* i *in the portfolio. (A trading strategy is also called a portfolio process, and* φ *itself is often called a portfolio.)*

The value of the portfolio φ *at time* t *is given by the expression*

$$\mathcal{V}_\varphi(t) = \sum_{i=0}^d \varphi_t^i X_t^i, \quad t \in [0,T].$$

The process $\mathcal{V}_\varphi(t)$ is called the value process *of the trading strategy φ.*

The gain process *$G_\varphi(t)$ is defined as*

$$G_\varphi(t) = \sum_{i=0}^{d} \int_0^t \varphi_s^i dX_s^i.$$

A trading strategy φ is called self-financing *if the wealth process $\mathcal{V}_\varphi(t)$ satisfies the equality*

$$\mathcal{V}_\varphi(t) = \mathcal{V}_\varphi(0) + G_\varphi(t), \quad t \in [0, T].$$

Let us recall that the market is arbitrage-free if there is no *arbitrage opportunity*, where an arbitrage opportunity is a self-financing trading strategy φ such that the wealth process $\mathcal{V}_\varphi(t)$ satisfies the conditions

$$\mathcal{V}_\varphi(0) = 0, \quad \mathbb{P}\big(\mathcal{V}_\varphi(T) \geq 0\big) = 1, \quad \mathbb{P}\big(\mathcal{V}_\varphi(T) > 0\big) > 0.$$

DEFINITION. 1.4 *Assume now that on $(\Omega, \mathcal{F})$ there exists a measure $\mathbb{P}^*$ equivalent to $\mathbb{P}$ such that the discounted price process $\beta_t X_t$ is a $\mathbb{P}^*$-martingale. The measure $\mathbb{P}^*$ is called a strong* martingale measure*, or a* risk-neutral measure*.*

*A self-financing trading strategy φ is called $\mathbb{P}^*$-*admissible *if the discounted gain process $\beta_t G_\varphi(t)$ is a $\mathbb{P}^*$-martingale.*

A contingent claim $g(X_T)$ is called attainable *if there exists an admissible trading strategy φ such that*

$$\mathcal{V}_\varphi(T) = g(X_T).$$

The trading strategy φ is then called a replicating strategy *for $g(X_T)$.*

Due to the no-arbitrage condition, the value of the replicating strategy φ_t defines the price $V(t)$ of the attainable contingent claim $g(X_T)$ at time t

$$V(t) = \mathcal{V}_\varphi(t).$$

For numerical applications, a more convenient pricing method is a method called the *risk-neutral pricing*.

THEOREM. 1.5 *Assume that on $(\Omega, \mathcal{F})$ there is a risk-neutral measure $\mathbb{P}^*$. Then the price of the $\mathbb{P}^*$-attainable contingent claim $g(X_T)$ at time $t \in [0, T]$ is*

$$V(t) = \beta_t^{-1} \mathbb{E}^*\big(\beta_T g(X_T) \big| \mathcal{F}_t\big),$$

where $\mathbb{E}^*$ denotes the expectation with respect to $\mathbb{P}^*$.

Remark. 1.2 *To simplify the notation, we assume that the measure $\mathbb{P}$ is already a strong martingale measure, and the discount process is constant $\beta_t = 1$. Under these simplifications, the valuation is reduced to the computation*

$$V(0) = \mathbb{E}\big(g(X_T)\big).$$

In some special cases we will consider variable β_t — a market risk-free *discount factor (typically a deterministic function of time).*

Remark. 1.3 *If the strong martingale measure $\mathbb{P}^*$ is unique then the market is complete. Hence, every contingent claim is attainable, and its price is uniquely defined.*

In incomplete markets there are many equivalent martingale measures, and non-attainable contingent claims have no uniquely defined prices. The prices which do not generate arbitrage opportunities usually cover a certain interval. To compute prices in that interval, in particular, to find endpoints of this interval, we can still use the formula from Theorem 1.5.

Black-Scholes model. The Black-Scholes model is a simple complete market model. There is only one risky asset in this model with the price process S_t and constant volatility σ. The discounting process is defined by a banking account with a constant interest rate r. As the market is complete, there is a unique martingale measure which we denote $\mathbb{P}$. The dynamic of the asset price in this martingale measure is given by the one-dimensional equation, which is a special case of equation (1.1)

$$dS_t = rS_t dt + \sigma S_t dW_t,$$

where r and σ are positive constants and W_t is a standard one-dimensional $\mathbb{P}$-Wiener process.

In this model the prices of European call and put options with strike K and maturity T are given by the celebrated Black-Scholes formula

$$V(t) = \kappa S_t \mathcal{N}(\kappa d_1) - \kappa K \mathrm{e}^{-r(T-t)} \mathcal{N}(\kappa d_2), \tag{1.2}$$

where $\kappa = 1$ corresponds to a call option, $\kappa = -1$, to a put option, and

$$d_1 = \frac{\ln(S_t/K) + (r + \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}}, \quad d_2 = d_1 - \sigma\sqrt{T-t}.$$

In the Black-Scholes model, closed-form formulas exist not only for call and put options and their combinations but also for several types of exotic options: binary options, barrier options, compound options, and some others. All these closed-form formulas are very useful in computational finance as they are benchmarks for numerical algorithms.

Local volatility models. In the Black-Scholes formula, the only unobservable parameter is the volatility σ. Since the formula can be inverted numerically, we can compute the volatility from the option prices observed in the financial market. The volatility computed in this way is called *implied volatility*. Under the assumptions of the Black-Scholes model, the implied volatility should be a constant function. In reality, we observe the *volatility smile*, a U-shaped function of the strike price K, which additionally is changing with the time to maturity T. The existence of the volatility smile indicates that assumptions of the Black-Scholes model should be relaxed. The simplest extension of the model is obtained by assuming that asset prices still follow a one-dimensional diffusion process but with a deterministic volatility function depending on the asset price and time. Such models are called *local volatility models*, and the price process in these models has the following dynamics

$$dS_t = b(t, S_t)dt + \sigma(t, S_t)dW_t. \tag{1.3}$$

If the functions $b(t, s)$ and $\sigma(t, s)$ in equation (1.3) are globally Lipschitz-continuous in s then, due to Theorem 1.2, there is a unique strong solution of (1.3). Unfortunately, in a number of popular local volatility models like the Constant Elasticity of Variance (CEV) model ($\sigma(t, s) = s^\beta$ with $0 < \beta \leq 1$), quadratic model ($\sigma(t, s) = \sigma_0 s^2$ with σ_0 a positive constant) or "limited" CEV (LCEV) model ($\sigma(t, s) = s \min(s^{\beta-1}, \epsilon^{\beta-1})$ with $0 < \beta \leq 1$ and ϵ a positive constant) the volatility function is not globally Lipschitz-continuous.

The following theorem gives the existence of a solution under relaxed assumptions.

THEOREM. 1.6 *Consider the d-dimensional stochastic differential equation*

$$dX_t^i = b_i(t, X_t)dt + \sum_{j=1}^{m} \sigma_i^j(t, X_t)dW_t^j, \quad i = 1, \ldots, d. \tag{1.4}$$

Assume that the coefficients of equation (1.4) *are locally Lipschitz-continuous*

$$|b_i(t, x) - b_i(t, y)| + |\sigma_i^j(t, x) - \sigma_i^j(t, y)| \leq K_n|x - y|, \quad |x|, |y| < n,$$

for $x, y \in D$ an open subset of $\mathbb{R}^d$ and $t \in [0, T]$. Then equation (1.4) *has a unique, strong solution up to an explosion time.*

If the coefficients of (1.4) *fulfill the linear growth condition*

$$|b_i(t, x)| + |\sigma_i^j(t, x)| \leq K(1 + |x|)$$

then the solution does not explode in a finite time.

The local Lipschitz condition is satisfied for the quadratic and LCEV models. From the above theorem, we conclude that for these models there exist unique, local solutions. But in the CEV model for $0 < \beta < 1$ the function $\sigma(t, s)$ is not locally Lipschitz-continuous. For this model the state $s = 0$ is absorbing and a unique solution exists only for $\frac{1}{2} \leq \beta \leq 1$ (cf. Andersen and Andreasen [2]).

Stochastic volatility models. Although local volatility models are easy to calibrate to market data and in many cases give closed-form expressions for option prices the dynamics produced by these models is not very realistic. In particular, the dynamics is not time invariant. To obtain a more realistic dynamics local volatility models have been expanded to *stochastic volatility models*. In these models the volatility is assumed to be a stochastic process dependent on a further exogenous parameter. We consider stochastic volatility models which are defines as a set of two correlated one-dimensional diffusion processes

$$\begin{aligned} dS_t &= b(t, S_t)dt + \sigma(Y_t)C(S_t)dW_t^S, \\ dY_t &= m(t, Y_t)dt + \nu(t, Y_t)dW_t^Y, \\ d\langle W^S, W^Y\rangle_t &= \rho dt. \end{aligned} \tag{1.5}$$

Formally, this model does not fit the market model described at the beginning of this Section as the volatility process Y_t is not a traded asset. Thus, the market model is incomplete, and there is no unique risk-neutral measure. On the other hand, the structure of system (1.5) is the same as equation (1.1) (if necessary after a de-correlation of the Wiener processes W^S and W^Y). As a result, Theorem 1.2 applies to many stochastic volatility models. But similarly, like for local volatility models, there are stochastic volatility models with coefficients not even locally Lipschitz-continuous. The existence of solutions for that models requires an individual analysis.

The risk-neutral option pricing described above can be easily approximated by several numerical methods. The most popular are Monte Carlo simulations and, thanks to the Feynman-Kac formula, numerical solutions of partial differential equations. We will analyze both of these approaches in detail in the subsequent chapters. We will also briefly describe binomial and trinomial trees, which are easy to implement and, in many cases, give quite accurate results.

1.2 Binomial trees

In this section, we present an application of binomial trees to pricing European style contingent claims written on a single risky asset whose dynamic in a risk-neutral measure is given by the Black-Scholes model

$$dS(t) = rS(t)dt + \sigma S(t)dW_t,\ S(0) = S_0,\ 0 \le t \le T. \tag{1.6}$$

We discretize the time interval $[0, T]$ and the process $S(t)$

$$\delta t = \frac{T}{N}, \quad t_n = n\delta t, \quad S_n = S(t_n).$$

The dynamic of S_n is given by the equation

$$S_{n+1} = Z_{n+1}S_n,$$

where Z_n is a sequence of independent random variables which can take for each n only two states (we assume $u > d$)

$$Z_n = \begin{cases} u, & \text{with probability } p, \\ d, & \text{with probability } 1 - p. \end{cases}$$

We assume a constant risk-free interest rate which gives the numéraire

$$B_n = \mathrm{e}^{r\, n\delta t}.$$

Then in risk-neutral measure

$$\mathbb{E}(S_{n+1}|S_n) = \mathrm{e}^{r\delta t}S_n. \tag{1.7}$$

This leads to the equation

$$pu + (1 - p)d = \mathrm{e}^{r\delta t},$$

which gives the risk-neutral probability

$$p = \frac{\mathrm{e}^{r\delta t} - d}{u - d}. \tag{1.8}$$

The condition $0 < p < 1$ is equivalent to

$$d < \mathrm{e}^{r\delta t} < u, \tag{1.9}$$

and guarantees the existence of a unique risk-neutral measure for the binomial model. The picture of the first few nodes of a binomial tree is seen on Fig. 1.1.

The pricing process of an option with payoff $g(S(T))$ can be described by the following steps:

1. Fix the input parameters: u, d, r, T and the number of time steps N.

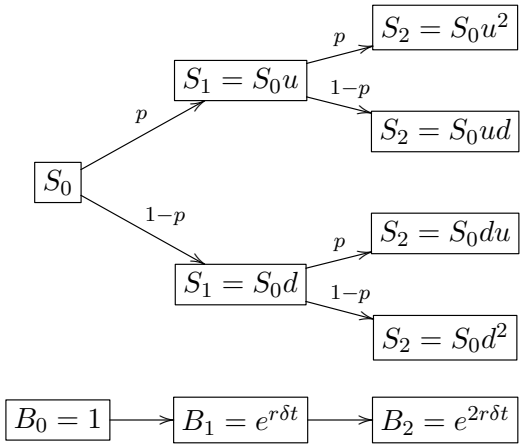

Figure 1.1: Binomial tree.

2. Compute the risk-neutral probability p.

3. Compute the payoff $g(S)$ in each node in period N.

4. For each n, starting from $N-1$ to 0, compute the payoff in each node using formula (1.7), the payoffs of the previous step, and the value of p.

5. The value in the root of the tree is the price of the option.

The above algorithm has a very high computational complexity as the number of nodes increases exponentially with the number of time steps. This algorithm can only be run with a small number of time steps, which can result in inaccurate pricing. To reduce the computational complexity, we have to limit the algorithm to the so-called *recombining binomial trees*, where we glue identical nodes into one node. A recombining binomial tree can be seen in Fig. 1.2.

Recombining trees can be used to compute prices of European and American contingent claims, but not for pricing path-dependent instruments (exotic).

Let S_{ji} denote the price after i time steps with j being the number of up moves. Since we deal with a recombining tree, then

$$S_{ji} = S_0 u^j d^{i-j}.$$

In a risk-neutral measure, we have

$$S_{ji} = \mathrm{e}^{-r\delta t}\Big(pS_{j+1,i+1} + (1-p)S_{j,i+1}\Big).$$

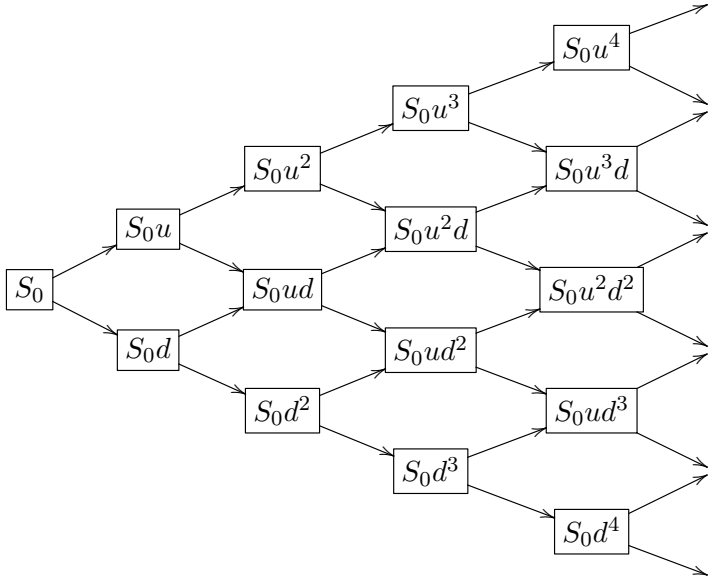

Figure 1.2: Recombining binomial tree.

For the payoff function $g(S(t))$, the price of a European option can be obtained by the sequence of iterates

$$\begin{aligned} V_{jN} &= g(S_{jN}), \qquad j = 0, 1, \ldots, N, \\ V_{ji} &= \mathrm{e}^{-r\delta t}\Big(pV_{j+1,i+1} + (1-p)V_{j,i+1}\Big), \quad i < N. \end{aligned}$$

For an American option we have

$$\begin{aligned} V_{jN} &= g(S_{jN}), \qquad j = 0, 1, \ldots, N, \\ V_{ji} &= \max\Big(g(S_{ji}), \mathrm{e}^{-r\delta t}\Big(pV_{j+1,i+1} + (1-p)V_{j,i+1}\Big)\Big), \quad i < N. \end{aligned}$$

To use this algorithm in practical computations, we have to calibrate the model to the market. The parameters r and S_0 are observed on the market, but the parameters u and d are only model idealizations and cannot be directly determined by the observation of market data. We turn to the Black-Scholes model (1.6) in which we know the asset prices

$$S(t) = S_0\mathrm{e}^{(r-\frac{1}{2}\sigma^2)t+\sigma W_t}, \qquad S_0 > 0.$$

Hence, the expected value of $S(t+\delta t)$ conditional on $S(t)$ is

$$\mathbb{E}\big(S(t+\delta t)|S(t)\big) = S(t)\mathrm{e}^{r\delta t},$$

and the conditional variance

$$\mathrm{Var}\big(S(t+\delta t)|S(t)\big) = S(t)^2\mathrm{e}^{2r\delta t}\Big(\mathrm{e}^{\sigma^2\delta t}-1\Big).$$

In the binomial tree the corresponding values are

$$\begin{aligned}
&\mathbb{E}\big(S(t+\delta t)|S(t)\big) = S(t)\big(pu+(1-p)d\big),\\
&\mathrm{Var}\big(S(t+\delta t)|S(t)\big) = \mathbb{E}\big(S^2(t+\delta t)|S(t)\big) - \Big(\mathbb{E}\big(S(t+\delta t)|S(t)\big)\Big)^2\\
&\qquad\qquad = p(S(t)u)^2 + (1-p)(S(t)d)^2 - \big(S(t)\big)^2\big(pu+(1-p)d\big)^2.
\end{aligned}$$

Comparing the right hand sides of the relevant equations we get

$$\begin{aligned}
&S(t)\big(pu+(1-p)d\big) = S(t)\mathrm{e}^{r\delta t},\\
&p(S(t)u)^2 + (1-p)(S(t)d)^2 - \big(S(t)\big)^2\big(pu+(1-p)d\big)^2\\
&\qquad\qquad = \big(S(t)\big)^2\mathrm{e}^{2r\delta t}\Big(\mathrm{e}^{\sigma^2\delta t}-1\Big).
\end{aligned}$$

After simplifications we have

$$\begin{aligned}
&pu+(1-p)d = \mathrm{e}^{r\delta t},\\
&pu^2+(1-p)d^2 = \mathrm{e}^{2r\delta t+\sigma^2\delta t}.
\end{aligned} \tag{1.10}$$

The first equation (1.10) gives the known expression of the risk-neutral probability

$$p = \frac{\mathrm{e}^{r\delta t}-d}{u-d},$$

the second equation is not sufficient to compute two parameters u and d. We need an additional relation. The choice of this relation gives rise to different binomial models. Here are some examples:

- the Cox-Ross-Rubinstein model (CRR) [13] in which we put $ud = 1$; this model is the industry standard,
- the Jarrow-Rudd model [27] in which we put $p = (1-p) = \frac{1}{2}$,

but there are many other possibilities used in practice.

We will now analyze the Cox-Ross-Rubinstein model.

THEOREM. 1.7 *Consider the CRR model for a fixed N with arbitrary u and d fulfilling conditions* (1.9) *and $ud = 1$, with p given by formula* (1.8)*. The price at time $t_n = n\delta t$ of a European call option with maturity T and strike K is*

$$V_N(t_n) = \mathrm{e}^{-r(N-n)\delta t} \sum_{j=0}^{N-n} \binom{N-n}{j} p^j (1-p)^{N-n-j} \big(S(t_n) u^j d^{N-n-j} - K\big)^+,$$

where the subscript N indicates the price computed in the model with N time steps.

Proof. Since $S(t_n) = S_0 \prod_{j=1}^n Z_j$ then using Theorem 1.5 we obtain

$$\begin{aligned}
V_N(t_n) &= \mathrm{e}^{-r(T-t_n)} \mathbb{E}\Big(\big(S_T - K\big)^+ \big| \mathcal{F}_{t_n} \Big) \\
&= \mathrm{e}^{-r(T-t_n)} \mathbb{E}\Big(\Big(S(t_n) \prod_{j=n+1}^{N} Z_j - K\Big)^+ \big| \mathcal{F}_{t_n} \Big) \\
&= \mathrm{e}^{-r(T-t_n)} \mathbb{E}\Big(\Big(S(t_n) \prod_{j=n+1}^{N} Z_j - K\Big)^+ \Big) \\
&= \mathrm{e}^{-r(N-n)\delta t} \sum_{j=0}^{N-n} \binom{N-n}{j} p^j (1-p)^{N-n-j} \big(S(t_n) u^j d^{N-n-j} - K\big)^+.
\end{aligned}$$

The above equalities follow from: the independence of Z_j of $\mathcal{F}_{t_n}$ for $j > n$, the $\mathcal{F}_{t_n}$-measurability of $S(t_n)$, and the nonnegativity of $(x-K)^+$. ■

Calibrating u and d in the Cox-Ross-Rubinstein model to the Black-Scholes model we obtain

$$\begin{aligned}
u &= \beta + \sqrt{\beta^2 - 1}, \\
d &= \beta - \sqrt{\beta^2 - 1},
\end{aligned}$$

where

$$\beta = \frac{1}{2}\Big(\mathrm{e}^{-r\delta t} + \mathrm{e}^{(r+\sigma^2)\delta t}\Big).$$

In practice, there is a tendency to use the simplified version of the CRR model

$$u = \mathrm{e}^{\sigma\sqrt{\delta t}}, \qquad d = \mathrm{e}^{-\sigma\sqrt{\delta t}}.$$

The simulations for the full and simplified models are, for large N, close to each other due to the convergence of the CRR model to the Black-Scholes model.

THEOREM. 1.8 *Consider a sequence of the CRR models starting from the same S_0 with N increasing to infinity. Let $\delta_N = \frac{T}{N}$. Define*

$$u_N = \mathrm{e}^{\sigma\sqrt{\delta_N}}, \quad d_N = \mathrm{e}^{-\sigma\sqrt{\delta_N}}, \quad p_N = \frac{\mathrm{e}^{r\delta_N} - \mathrm{e}^{-\sigma\sqrt{\delta_N}}}{\mathrm{e}^{\sigma\sqrt{\delta_N}} - \mathrm{e}^{-\sigma\sqrt{\delta_N}}}.$$

Let $V_N(0)$ denote the price at time $t = 0$ of a European call option with maturity T and strike K and the parameters u_N, d_N, p_N defined above.

When $V_{BS}(t)$ denotes the Black-Scholes price of that option at time t, then

$$\lim_{N\to\infty} V_N(0) = V_{BS}(0).$$

Proof. Let $S_n^N = S(n\delta_N)$ denote the asset prices in the model for a fixed N. Denote $Z_n^N = \frac{S_n^N}{S_{n-1}^N}$, then

$$S_n^N = S_0 \prod_{j=1}^{n} Z_j^N.$$

Since $\mathbb{P}(Z_j^N = u_N) = p_N$, $\mathbb{P}(Z_j^N = d_N) = 1 - p_N$, and $N\delta_N = T$, we obtain

$$S_N^N = S^N(T) = S_0 \exp\Big(\sigma\sqrt{\delta_N} \sum_{j=1}^{N} R_j^N\Big),$$

where $R_j^N = \frac{\ln Z_j^N}{\sigma\sqrt{\delta_N}}$.

R_j^N are independent (since Z_j^N are independent) and with the identical distribution

$$\mathbb{P}(R_j^N = 1) = p_N, \qquad \mathbb{P}(R_j^N = -1) = 1 - p_N, \text{ for } j = 1, \ldots, N.$$

Expanding p_N into Taylor's series, we obtain

$$p_N = \frac{1}{2} + \frac{r - \frac{1}{2}\sigma^2}{2\sigma}\sqrt{\delta_N} + O(N^{-1}).$$

Then for each j

$$\mathbb{E}\big(\sigma\sqrt{\delta_N} R_j^N\big) = \big(r - \tfrac{1}{2}\sigma^2\big)\frac{T}{N} + o(N^{-1})$$

and

$$\mathrm{Var}\big(\sigma\sqrt{\delta_N} R_j^N\big) = \sigma^2 \frac{T}{N} + o(N^{-1}).$$

Hence by the central limit theorem, we obtain

$$\sigma\sqrt{\frac{T}{N}}\sum_{j=1}^{N} R_j^N \to \mathcal{N}\big((r-\tfrac{1}{2}\sigma^2)T, \sigma^2 T\big) \text{ in distribution as } N\to\infty$$

and

$$S^N(T) \to S_0 \exp\big((r-\tfrac{1}{2}\sigma^2)T + \sigma\sqrt{T}Z\big) \text{ in distribution as } N\to\infty,$$

where $Z \sim \mathcal{N}(0,1)$.

Since

$$V_N(0) = \mathrm{e}^{-rT}\mathbb{E}\big((S^N(T)-K)^+\big),$$

then $V_N(0)$ converges to

$$\begin{aligned} V(0) &= \mathrm{e}^{-rT}\mathbb{E}\Big(\big(S_0\exp\big((r-\tfrac{1}{2}\sigma^2)T+\sigma\sqrt{T}Z\big)-K\big)^+\Big) \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\big(S_0\exp\big(-\tfrac{1}{2}\sigma^2 T+\sigma\sqrt{T}x\big)-\mathrm{e}^{-rT}K\big)^+\mathrm{e}^{-\frac{1}{2}x^2}dx. \end{aligned}$$

Let

$$\gamma = \frac{\ln(K/S_0)+(\frac{1}{2}\sigma^2-r)T}{\sigma\sqrt{T}}.$$

The integrand is non-zero only on the interval (γ,∞) as

$$\sigma\sqrt{T}x - \tfrac{1}{2}\sigma^2 T > \ln(K/S_0) - rT$$

only on this interval. Then we have

$$\begin{aligned} V(0) &= \frac{1}{\sqrt{2\pi}}S_0\int_{\gamma}^{\infty}\mathrm{e}^{-\frac{1}{2}\sigma^2 T}\mathrm{e}^{\sigma\sqrt{T}x-\frac{1}{2}x^2}dx - K\mathrm{e}^{-rT}\big(1-\Phi(\gamma)\big) \\ &= \frac{1}{\sqrt{2\pi}}S_0\int_{\gamma}^{\infty}\mathrm{e}^{-\frac{1}{2}(x-\sigma\sqrt{T})^2}dx - K\mathrm{e}^{-rT}\big(1-\Phi(\gamma)\big) \\ &= S_0\big(1-\Phi(\gamma-\sigma\sqrt{T})\big) - K\mathrm{e}^{-rT}\big(1-\Phi(\gamma)\big), \end{aligned}$$

where Φ is the cumulative normal distribution function.

By symmetry $\big(1-\Phi(\gamma)\big) = \Phi(-\gamma)$ we obtain

$$\begin{aligned} -\gamma+\sigma\sqrt{T} &= \frac{\ln(S_0/K)+(r+\frac{1}{2}\sigma^2)T}{\sigma\sqrt{T}} \equiv d_+, \\ -\gamma &= \frac{\ln(S_0/K)+(r-\frac{1}{2}\sigma^2)T}{\sigma\sqrt{T}} \equiv d_-. \end{aligned}$$

Hence

$$V(0) = S_0\Phi(d_+) - \mathrm{e}^{-rT}K\Phi(d_-)$$

which is the Black-Scholes formula for call options. ■

In the Jarrow-Rudd model we have

$$u = \mathrm{e}^{r\delta t}\big(1 + \sqrt{\mathrm{e}^{\sigma^2\delta t} - 1}\big),$$
$$d = \mathrm{e}^{r\delta t}\big(1 - \sqrt{\mathrm{e}^{\sigma^2\delta t} - 1}\big).$$

Very often the following simplified version of these parameters is used:

$$u = \mathrm{e}^{(r-\sigma^2/2)\delta t + \sigma\sqrt{\delta t}},$$
$$d = \mathrm{e}^{(r-\sigma^2/2)\delta t - \sigma\sqrt{\delta t}}.$$

For a general binomial model, we can assume without loss of generality

$$ud = \mathrm{e}^{2\nu\delta t},$$

for a scalar parameter ν. This leads to the following simplified version of parameters u and d (the exact expressions are very complicated):

$$u = \mathrm{e}^{\nu\delta t + \sigma\sqrt{\delta t}},$$
$$d = \mathrm{e}^{\nu\delta t - \sigma\sqrt{\delta t}},$$
$$p = \frac{1}{2} + \frac{1}{2}\Big(\frac{\mu - \nu}{\sigma}\Big)\sqrt{\delta t}, \quad \text{where } \mu = r - \sigma^2/2.$$

The value $\nu = 0$ corresponds to the CRR model (the tree is symmetric) and $\nu = \mu$, to the Jarrow-Rudd model (there is a deterministic drift $\mathrm{e}^{\mu\delta t}$).

Binomial models are the simplest lattice models used in financial computations. But trees with a larger number of states are also in use. We present an implementation of trinomial trees. We assume, like for binomial trees, that

$$S_{n+1} = Z_{n+1}S_n.$$

For trinomial trees, Z_n is a sequence of independent random variables which, for each n, take 3 states

$$Z_n = \begin{cases} u, & \text{with probability } p_u, \\ m, & \text{with probability } p_m, \\ d, & \text{with probability } p_d. \end{cases}$$

Our goal is to find p_u, p_m and p_d which define a risk-neutral measure. We have the obvious condition $p_u + p_m + p_d = 1$. In addition, to obtain a recombining tree we need $ud = m^2$, where we assume $u > m > d$. Computing $\mathbb{E}(S_{n+1}|S_n)$ in risk-neutral measure we obtain

$$p_u u + p_m m + p_d d = \mathrm{e}^{r\delta t}.$$

The above 3 equations are not sufficient to determine uniquely a risk-neutral measure. Hence there are many risk-neutral measures for trinomial trees, and the market is incomplete. Thus, not all contingent claims can be uniquely priced in this model. (Note the difference with binomial models where a risk-neutral measure is uniquely defined by (1.8) independently from the model specification.)

The process of calibration for trinomial trees is similar to the process for binomial trees. The comparison of conditional expectations and variances between the trinomial model and the Black-Scholes model gives the equations

$$\begin{aligned} p_u u + p_m m + p_d d &= \mathrm{e}^{r\delta t}, \\ p_u u^2 + p_m m^2 + p_d d^2 &= \mathrm{e}^{(2r+\sigma^2)\delta t}. \end{aligned} \tag{1.11}$$

Together with equations $p_u + p_m + p_d = 1$ and $ud = m^2$ we have 4 equations for 6 unknowns. To make the problem solvable we assume

$$Z_n = \begin{cases} \mathrm{e}^{\mu\delta t + \lambda\sigma\sqrt{\delta t}}, & \text{with probability } p_u, \\ \mathrm{e}^{\mu\delta t}, & \text{with probability } p_m, \\ \mathrm{e}^{\mu\delta t - \lambda\sigma\sqrt{\delta t}}, & \text{with probability } p_d. \end{cases}$$

Computing u, m and d by this approximation and inserting $p_m = 1 - p_u - p_d$ in (1.11), we obtain

$$\begin{aligned} p_u(U-1) + p_d(D-1) &= F - 1, \\ p_u(U^2-1) + p_d(D^2-1) &= H - 1, \end{aligned}$$

where $U = \mathrm{e}^{\lambda\sigma\sqrt{\delta t}}$, $D = \mathrm{e}^{-\lambda\sigma\sqrt{\delta t}}$, $F = \mathrm{e}^{(r-\mu)\delta t}$ and $H = \mathrm{e}^{(2r+\sigma^2-2\mu)\delta t}$. Solving the above equations we get

$$\begin{aligned} p_u &= \frac{H - (D+1)F + D}{(U-D)(U-1)}, \\ p_m &= \frac{-H + (U+D)F - UD}{(1-D)(U-1)}, \\ p_d &= \frac{H - (U+1)F + U}{(U-D)(1-D)}. \end{aligned} \tag{1.12}$$

This model has a built-in deterministic drift $\mathrm{e}^{\mu\delta t}$. To obtain a symmetric trinomial tree we assume $\mu = 0$. That symmetric model with

$$Z_n = \begin{cases} \mathrm{e}^{\lambda\sigma\sqrt{\delta t}}, & \text{with probability } p_u, \\ 1, & \text{with probability } p_m, \\ \mathrm{e}^{-\lambda\sigma\sqrt{\delta t}}, & \text{with probability } p_d. \end{cases} \tag{1.13}$$

can be considered as a trinomial version of the CRR model. Expanding all terms in equations (1.12) in power series of $\sqrt{\delta t}$ we obtain after tedious computations the following expressions in the first-order approximation

$$\begin{aligned} p_u &= \frac{1}{2\lambda^2} + \frac{(r - \sigma^2/2)\sqrt{\delta t}}{2\lambda\sigma}, \\ p_m &= 1 - \frac{1}{\lambda^2}, \\ p_d &= \frac{1}{2\lambda^2} - \frac{(r - \sigma^2/2)\sqrt{\delta t}}{2\lambda\sigma}. \end{aligned} \tag{1.14}$$

Kamrad and Ritchken have obtained a similar result under the additional assumptions that $\mathbb{E}(\ln Z_n) = (r - \sigma^2/2)\delta t$ and $\mathrm{Var}(\ln Z_n) = \sigma^2\delta t$ which corresponds to a discrete version of Black-Scholes. That assumptions greatly simplify the computations and we obtain the following match of the first two moments of $\ln Z_n$

$$\begin{aligned} \lambda\sigma\sqrt{\delta t}(p_u - p_d) &= (r - \tfrac{1}{2}\sigma^2)\delta t, \\ \lambda^2\sigma^2\delta t(p_u + p_d) &= \sigma^2\delta t. \end{aligned}$$

Solving that system of equations, we obtain (1.14) as exact solutions.

Thus for the symmetric trinomial tree, we know the parameters of the model as the functions of λ. Let us notice that to obtain the non-negative probabilities, we have to take $\lambda \geq 1$. But, we also have the stability problem of the model. $\lambda = 1$ is on the edge of the stability region. To get a stable model, we have to choose λ substantially larger than 1. The choice of λ influences the rate of convergence of option prices computed in the model to the Black-Scholes prices. Kamrad and Ritchken have shown that the value of λ that produces the best convergence rate is

$$\lambda = \sqrt{3/2}.$$

1.3 Monte Carlo methods

We describe the idea behind Monte Carlo methods in a simple example: the computation of expectation $\mathbb{E}(g(X_T))$ where X_T is a random variable (the state at time

T of a stochastic process X_t). Let us assume that the distribution $\phi_X(x)$ of the random variable X_T is known. Then

$$\mathbb{E}\big(g(X_T)\big) = \int_{\mathbb{R}^d} g(x)\phi_X(x)dx = \int_{[0,1]^d} f(x)dx,$$

where $f(x)$ is obtained by a change of variables $\mathbb{R}^d \to [0,1]^d$.

The function $f(x)$ is in many cases too complicated to obtain an analytic formula for the integral $\int_{[0,1]^d} f(x)dx$. We can compute that integral by the following Monte Carlo algorithm: we sample N points x_i distributed "uniformly" in the cube $[0,1]^d$ and use the approximation

$$\int_{[0,1]^d} f(x)dx \simeq \frac{1}{N}\sum_{i=1}^{N} f(x_i). \tag{1.15}$$

This algorithm is similar to a deterministic quadrature. The difference lies in error estimates: a deterministic quadrature of order k approximates the integral with the error $O(N^{-k/d})$ and the Monte Carlo approximation, using randomly sampled points, gives the error of order $O(N^{-1/2})$ independent of the dimension d (see comments below).

In practice, we omit computation of $f(x)$ and sample points x_i from the distribution of X_T

$$\mathbb{E}(g(X_T)) \simeq \frac{1}{N}\sum_{i=1}^{N} g(x_i). \tag{1.16}$$

The foundation of Monte Carlo methods consists of two results: the strong law of large numbers and the central limit theorem.

THEOREM. 1.9 (Strong law of large numbers) *Let $X_1, X_2, \ldots$ be a sequence of square integrable, identically distributed, independent random variables with $\mathbb{E}(X_i) = \mu < \infty$. Let*

$$Y_n = \frac{X_1 + X_2 + \cdots + X_n}{n}, \qquad n = 1, 2, \ldots.$$

Then

$$\lim_{n\to\infty} Y_n = \mu, \text{ a.s.}$$

The strong law of large numbers implies that the right hand side averages in equations (1.15) and (1.16) converge to the integrals in the left hand sides a.s.

THEOREM. 1.10 (Central limit theorem) *Let $X_1, X_2, \ldots$ be a sequence of square integrable, identically distributed, independent random variables with expectations $\mathbb{E}(X_i) = \mu < \infty$ and variances $\mathrm{Var}(X_i) = \sigma^2$. Define the sequence*

$$Z_n = \frac{(X_1 - \mu) + (X_2 - \mu) + \cdots + (X_n - \mu)}{\sqrt{n}}.$$

Then

$$Z_n \to \mathcal{N}(0, \sigma^2) \text{ in distribution.}$$

The central limit theorem implies that if x_i, $i = 1, \ldots, N$, is a sample drawn independently from the distribution of X_T, then the estimator

$$\hat{g}_N = \frac{1}{N} \sum_{i=1}^{N} g(x_i)$$

converges in distribution to $\mathcal{N}(\mu, \frac{\sigma^2}{N})$ where $\mu = \mathbb{E}(g(X_T))$. This gives the mentioned earlier estimate of the Monte Carlo error $\mathrm{Var}(\hat{g}_N) = O(N^{-1})$.

In Monte Carlo computations we need independent samples x_i, $i = 1, \ldots, N$, from a given distribution. To generate such samples one needs random number generators with sufficiently good properties. We will describe such generators in Chapter 2.

1.4 Methods of partial differential equations

Consider a stochastic process given by the equation

$$dX_s = b(s, X_s)ds + \sigma(s, X_s)dW_s, \quad X_t = x. \tag{1.17}$$

Let us assume that the coefficients in this equation fulfill the conditions of Theorem 1.2. Hence, the equation possesses a unique strong solution which we denote $X_s^{t,x}$ to indicate the dependence on initial data $X_t = x$.

DEFINITION. 1.11 *Let $X_s^{t,x}$ be a solution of* (1.17) *with the coefficients b and σ independent of s. The (infinitesimal) generator $\mathcal{A}$ of X_s is defined by*

$$\mathcal{A}u(x) = \lim_{s \to t} \frac{\mathbb{E}\big(u(X_s^{t,x})\big) - u(x)}{s - t}, \quad x \in \mathbb{R}^d.$$

The set of functions u: $\mathbb{R}^d \to \mathbb{R}$ for which the limit exists for all $x \in \mathbb{R}^d$, is denoted by $\mathcal{D}_\mathcal{A}$ and is called the domain of $\mathcal{A}$.

For a function $u \in \mathcal{D}_{\mathcal{A}}$

$$\mathcal{A}u(x) = \sum_{i,j=1}^{d} a_{ij}(x)\frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^{d} b_i(x)\frac{\partial u}{\partial x_i}, \quad \textit{where } a_{ij} = \frac{1}{2}\sum_{k=1}^{m} \sigma_i^k \sigma_j^k.$$

When the coefficients in (1.17) *are time dependent* $b = b(t,x)$ *and* $\sigma = \sigma(t,x)$*, we have the family of operators* $\mathcal{A}^t$

$$\mathcal{A}^t u(t,x) = \sum_{i,j=1}^{d} a_{ij}(t,x)\frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^{d} b_i(t,x)\frac{\partial u}{\partial x_i}.$$

The computation of expectation $\mathbb{E}\big(g(X_T^{t,x})\big)$ can be reduced to a solution of a partial differential equation due to the following theorem.

THEOREM. 1.12 (Feynman-Kac) *Consider the Cauchy problem*

$$\begin{aligned} &-\tfrac{\partial v}{\partial t} - \mathcal{A}^t v + kv - f = 0, \quad t \in [0,T), x \in \mathbb{R}^d, \\ &v(T,x) = g(x), \quad x \in \mathbb{R}^d, \end{aligned} \tag{1.18}$$

where $\mathcal{A}^t$ *is given in Definition 1.11 and is assumed to be uniformly elliptic, i.e.*

$$\exists \delta > 0, \quad \sum_{i,j=1}^{d} a_{ij}(t,x)\xi_i\xi_j \geq \delta|\xi|^2, \quad \forall t \in [0,T],\ x \in \mathbb{R}^d,\ \xi \in \mathbb{R}^d \setminus \{0\}.$$

Assume that the functions $k\colon\ [0,T] \times \mathbb{R}^d \to [0,\infty)$*,* $f\colon\ [0,T] \times \mathbb{R}^d \to \mathbb{R}$ *and* $g\colon\ \mathbb{R}^d \to \mathbb{R}$ *are continuous and satisfy (with some constants* C *and* $q \geq 2$*)*

$$\begin{aligned} &|g(x)| \leq C(1+|x|^q),\ \textit{or}\ \ g(x) \geq 0, \quad \forall x \in \mathbb{R}^d, \\ &|f(t,x)| \leq C(1+|x|^q),\ \textit{or}\ \ f(t,x) \geq 0, \quad t \in [0,T], \forall x \in \mathbb{R}^d. \end{aligned} \tag{1.19}$$

Let $v(t,x)$ *be a solution of* (1.18) *which is continuous in* $[0,T] \times \mathbb{R}^d$ *and of class* $C^{1,2}\big([0,T)\times\mathbb{R}^d\big)$ *with the polynomial growth condition (for some constants* $C > 0$ *and* $p \geq 2$*)*

$$\sup_{t\in[0,T]} |v(t,x)| \leq C(1+|x|^p), \quad x \in \mathbb{R}^d.$$

Then $v(t,x)$ *is a unique solution of* (1.18) *and admits the stochastic representation*

$$\begin{aligned} v(t,x) =& \mathbb{E}\bigg(g(X_T^{t,x})\exp\Big(-\int_t^T k(r,X_r^{t,x})dr\Big) \\ &+ \int_t^T f(s,X_s^{t,x})\exp\Big(-\int_t^s k(r,X_r^{t,x})dr\Big)ds\bigg). \end{aligned}$$

Proof. To simplify the proof we take $dX_s = dW_s$ with $X_t = x$, and $k = k(x)$. Then $\mathcal{A}^t u = \frac{1}{2}\Delta u$. Consider the expression

$$v(s, X_s) \exp\Big(-\int_t^s k(X_u)du\Big).$$

We obtain from Itô's rule and equation (1.18)

$$\begin{aligned} d_s\Big(v(s,X_s)\exp\Big(-\int_t^s k(X_u)du\Big)\Big) &= \exp\Big(-\int_t^s k(X_u)du\Big) \\ &\times \left(\frac{\partial v}{\partial s}ds - k(X_s)vds + \sum_{i=1}^d \frac{\partial}{\partial x_i}v(s,X_s)dW_s^i + \frac{1}{2}\Delta v\, ds\right) \\ &= \exp\Big(-\int_t^s k(X_u)du\Big)\Big(-f(s,X_s)ds + \sum_{i=1}^d \frac{\partial}{\partial x_i}v(s,X_s)dW_s^i\Big). \end{aligned}$$

Let $S_n = \inf\{s \geq t\colon\ |X_s| \geq n\}$. We integrate on $[t, T \wedge S_n]$ and compute the expectation. Since the resulting stochastic integral has expectation zero, we obtain

$$\begin{aligned} v(t,x) =& \mathbb{E}\left(\int_t^{T\wedge S_n} f(s,X_s)\exp\Big(-\int_t^s k(X_u)du\Big)ds\right) \\ &+ \mathbb{E}\left(v(S_n, X_{S_n})\exp\Big(-\int_t^{S_n} k(X_u)du\Big)\mathbb{1}_{\{S_n\leq T\}}\right) \\ &+ \mathbb{E}\left(v(T,X_T)\exp\Big(-\int_t^T k(X_u)du\Big)\mathbb{1}_{\{S_n > T\}}\right). \end{aligned}$$

By dominated convergence, the first term in the above sum converges to

$$\mathbb{E}\left(\int_t^T f(s,X_s)\exp\Big(-\int_t^s k(X_u)du\Big)ds\right).$$

For the second term we have

$$\begin{aligned} &\mathbb{E}\left(v(S_n, X_{S_n})\exp\Big(-\int_t^{S_n} k(X_u)du\Big)\mathbb{1}_{\{S_n\leq T\}}\right) \\ &\quad\leq \mathbb{E}\Big(v(S_n, X_{S_n})\mathbb{1}_{\{S_n\leq T\}}\Big) \leq C(1+n^p)\mathbb{P}^x(S_n \leq T). \end{aligned}$$

By the Chebyshev inequality and Theorem 1.2 we have the estimate

$$\begin{aligned} \mathbb{P}^x(S_n \leq T) &= \mathbb{P}^x\Big(\sup_{0\leq t\leq T}|X_t| \geq n\Big) \\ &\leq n^{-k}\,\mathbb{E}^x\Big(\sup_{0\leq t\leq T}|X_t|^k\Big) \leq Cn^{-k}\big(1+|x|^k\big). \end{aligned}$$

Taking $k > p$, we see that the second term converges to zero.

By the dominated convergence theorem, the third term converges to

$$\mathbb{E}\Big(v(T, X_T)\exp\Big(-\int_t^T k(X_u)du\Big)\Big).$$

Since $v(T, X_T) = g(X_T)$ this ends the proof. ■

The following theorem follows from a probabilistic construction of solutions to equation (1.18). It can be regarded as the counterpart to the Feynman-Kac theorem.

THEOREM. 1.13 *Let $X_s^{t,x}$ be a strong solution of* (1.17)*. We assume that the coefficients $b(s,x)$ and $\sigma(s,x)$ of that equation fulfill the assumptions of Theorem 1.2. In addition, they are twice continuously differentiable with respect to x for any $s \in [0,T]$ and all their partial derivatives with respect to x up to the second-order are bounded.*

Let $k\colon\ [0,T]\times\mathbb{R}^d \to [0,\infty)$, $f\colon\ [0,T]\times\mathbb{R}^d \to \mathbb{R}$ and $g\colon\ \mathbb{R}^d \to \mathbb{R}$ be Borel functions that are twice continuously differentiable with respect to x for any $t \in [0,T]$. Assume that absolute values of these functions and all their partial derivatives with respect to x up to the second-order are bounded by $C(1+|x|^q)$ (with some constants $C > 0$ and $q \geq 2$). Then

$$\begin{aligned} u(t,x) =& \mathbb{E}\bigg(g(X_T^{t,x})\exp\Big(-\int_t^T k(r, X_r^{t,x})dr\Big) \\ &+ \int_t^T f(s, X_s^{t,x})\exp\Big(-\int_t^s k(r, X_r^{t,x})dr\Big)ds\bigg) \end{aligned}$$

is a well defined function in $[0,T]\times\mathbb{R}^d$ which is twice differentiable with respect to x continuously in (t,x) and continuously differentiable with respect to t in $[0,T]\times\mathbb{R}^d$ with the growth estimate (for some constants $C > 0$ and $p \geq 2$)

$$\sup_{0\leq t\leq T}|u(t,x)| \leq C(1+|x|^p), \quad x \in \mathbb{R}^d.$$

The function $u(t,x)$ is a solution of equation (1.18) *which fulfills the terminal condition $\lim_{t\to T} u(t,x) = g(x)$.*

Remark. 1.4 *Let in addition to the assumptions of Theorem 1.13 functions σ_i^j and b_i be n times continuously differentiable with respect to x for any $s \in [0,T]$ ($n \geq 2$) and all partial derivatives of these functions with respect to x up to order n be bounded. We require also that k, f and g are n times continuously differentiable with respect to x for any $t \in [0,T]$ and absolute values of these functions and all*

their partial derivatives with respect to x up to order n are bounded by $C(1+|x|^q)$. Then $u(t,x)$ defined by Theorem 1.13 is n times differentiable with respect to x continuously in (t,x) and continuously differentiable with respect to t in $[0,T]\times\mathbb{R}^d$ with the absolute values of all its partial derivatives with respect to x up to order n bounded by $C(1+|x|^p)$ for some constants $C>0$ and $p\geq 2$.

The proof of Theorem 1.13 for sufficiently regular coefficients follows from theorems of Chapter 5 and Sobolev's inequalities, but under the assumed regularity, the proof requires probabilistic solutions of PDEs (see the book by Krylov[32]). Remark 1.4 can be proved following the line of the proof of Theorem V.7.4 in [32].

The following corollary from the Feynman-Kac theorem is essential in the derivation of the Black-Scholes equation.

COROLLARY. 1.14 *If, in the Black-Scholes model, the price of underlying in a risk-neutral measure fulfills the equation*

$$dS_t = rS_t dt + \sigma S_t dW_t,$$

then the price of the contingent claim with payoff $g(S_T)$ is

$$V_g(t) = \mathbb{E}\Big(\mathrm{e}^{-r(T-t)}g(S_T)\big|S_t = x\Big).$$

By the Feynman-Kac theorem $G(t,x) = \mathbb{E}\big(g(S_T)|S_t = x\big)$ fulfills the equation

$$\frac{\partial G}{\partial t} + rx\frac{\partial G}{\partial x} + \frac{1}{2}\sigma^2 x^2 \frac{\partial^2 G}{\partial x^2} = 0.$$

Since $F(t,x) := V_g(t) = \mathrm{e}^{-r(T-t)}G(t,x)$, we have

$$\frac{\partial F}{\partial t} + rx\frac{\partial F}{\partial x} + \frac{1}{2}\sigma^2 x^2 \frac{\partial^2 F}{\partial x^2} - rF = 0, \quad F(T,x) = g(x),$$

which is the Black-Scholes equation.

By the Feynman-Kac theorem and the above corollary, computation of $\mathbb{E}(g(X_T))$ can be reduced to the construction of a smooth solution of a certain partial differential equation. The existence of such solutions and their numerical approximations will be discussed in Chapters 5, 6, and 7.

Chapter 2

Random number generators

To generate truly random numbers, we have to use a physical device with random behavior. Such generators are not used in modern digital computers (there are some exceptions such as generating seeds for pseudo-random generators). "Random numbers" generated by digital computers are obtained by deterministic algorithms. We require only that a sequence of such numbers mimics sufficiently well the statistical properties of true random numbers. To indicate a deterministic character of generated sequences we call them *pseudo-random* numbers. The majority of generators used in practice are generators of uniform deviates on $[0, 1]$. Generators of other distributions are obtained by a suitable transformation of uniform distributions.

2.1 Generators of uniform deviates

Linear congruential generators are the simplest generators of pseudo-random numbers.

DEFINITION. 2.1 *A linear congruential generator is a recurrence of the form:*

1. *Choose* x_0 *(seed);*
2. *Compute* $x_i = (ax_{i-1} + b) \bmod M$ *for some integers* a*,* b *and* M*;*
3. *Compute* $u_i = \frac{x_i}{M}$.

COROLLARY. 2.2 *The properties of sequences generated by linear congruential generators:*

1. $x_i \in \{0, 1, \ldots, M-1\}$*;*

2. *The sequence x_i is periodic with period not greater than M.*

But a wrong choice of a, b, or M can result in a period much smaller than M !

Sequences $\{u_i\}$ obtained by linear congruential generators should be independent samples of a random variable with uniform distribution. We can investigate this independence by analyzing n-tuples $(u_i, u_{i+1}, \ldots, u_{i+n-1})$.

Example. 2.3 *Let us consider two-dimensional vectors. We have*

$$x_i = (ax_{i-1} + b) \bmod M = ax_{i-1} + b - kM, \text{ for } kM \le ax_{i-1} + b < (k+1)M.$$

For arbitrary z_0, z_1, we obtain

$$\begin{aligned} &z_0 x_{i-1} + z_1 x_i = z_0 x_{i-1} + z_1(ax_{i-1} + b - kM) \\ &= x_{i-1}(z_0 + az_1) + z_1 b - z_1 kM = M\Big(x_{i-1}\frac{z_0 + az_1}{M} - z_1 k\Big) + z_1 b. \end{aligned}$$

Denoting the quantity in parenthesis by c we obtain the equation

$$z_0 u_{i-1} + z_1 u_i = c + z_1 b M^{-1}. \tag{2.1}$$

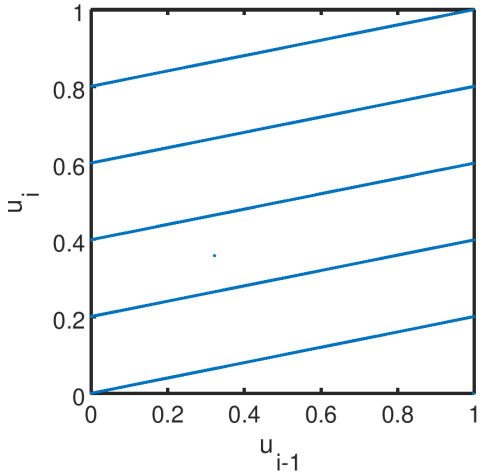

Figure 2.1: The points (u_{i-1}, u_i) for $a = 1229$, $b = 1$, $M = 2048$.

That is the equation of a straight line. Choosing integers z_0 and z_1 and requiring $z_0 + az_1 = 0 \bmod M$ we obtain c which is integer. That observation will

be used in the subsequent analysis. The straight line given by equation (2.1) *depends on* i*, since* c *is a function of* x_{i-1}*. We expect sufficiently many of such lines covering densely the square* $[0,1)^2$*. But that expectation appears to be wrong in many practical situations. For many* a *and* M *the points* (u_{i-1}, u_i) *lie on very few straight lines. As an example consider the linear congruential generator with* $a = 1229$*,* $b = 1$*,* $M = 2048$ *and* $x_0 = 1$*. Choosing* $z_0 = -1$*, and* $z_1 = 5$*, we get* $-1 + 1229 \cdot 5 = 6144 = 3 \cdot 2048 = 0 \bmod 2048$ *which gives an integer* c*. By equation* (2.1) *and inequality* $0 \leq u_i < 1$*, the value of* c *can only be a number from* $\{-1, 0, 1, 2, 3, 4\}$*. Hence, there are only 6 straight lines on which vectors* (u_{i-1}, u_i) *lie. Figure 2.1 shows the picture of such vectors for* 5000 *simulations. In that picture we observe only 5 lines. The explanation comes from the fact that* $c = -1$ *corresponds to the straight line which is in the lower right corner of the figure and is not visible.*

A better choice of a*,* b *and* M *can produce a better picture. Let us take* $a = 2^{16} + 3$*,* $b = 0$*,* $M = 2^{31}$*. As Figure 2.2 suggests, we obtained a deceptively good congruential generator.*

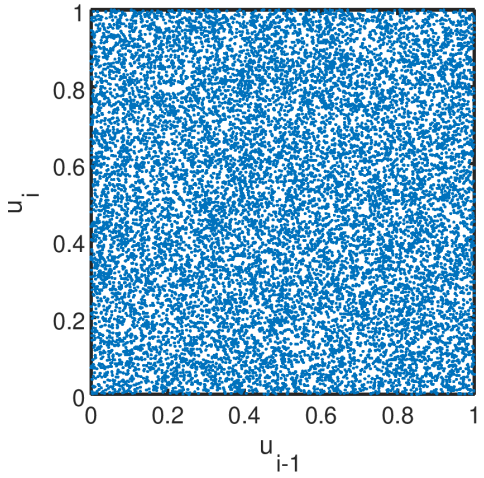

Figure 2.2: The points (u_{i-1}, u_i) for $a = 2^{16} + 3$, $b = 0$, $M = 2^{31}$.

This positive picture breaks down with 3-dimensional vectors (u_{i-2}, u_{i-1}, u_i)*. Figure 2.3 shows that 3-dimensional vectors are concentrated on a small number of planes in the 3-dimensional cube* $[0,1)^3$ *(analysis similar as for the previous generator reveals that there are 15 such planes).*

The bad behavior of linear congruential generators observed in the above example is not exceptional. The following theorem due to Marsaglia [38], which we

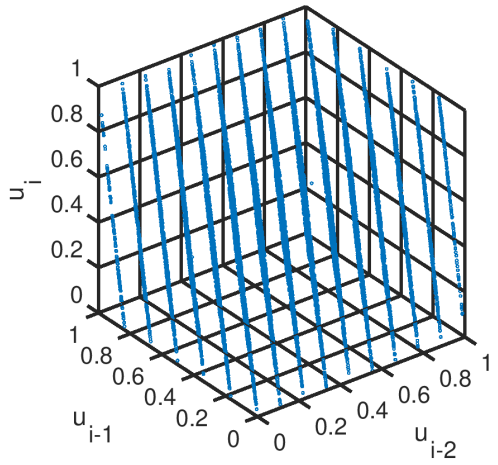

Figure 2.3: The points (u_{i-2}, u_{i-1}, u_i) for $a = 2^{16} + 3$, $b = 0$, $M = 2^{31}$.

present without proof, explains that such behavior is typical.

THEOREM. 2.4 *For each linear congruential generator there is an $n \ll M$ such that n-tuples $(u_i, u_{i+1}, \ldots, u_{i+n-1})$ lie on a relatively low number of hyperplanes in $\mathbb{R}^n$.*

To avoid the effect of correlation described above, one uses more sophisticated generators, for example – the generalized congruential generators

$$x_i = (a_1 x_{i-1} + \cdots + a_k x_{i-k}) \bmod M.$$

L'Ecuyer [34] recommends designing mixed generators from several generalized congruential generators. Taking J generalized generators

$$x_{j,i} = (a_{j,1} x_{j,i-1} + \cdots + a_{j,k} x_{j,i-k}) \bmod M_j, \quad j = 1, \ldots, J,$$

we can construct the mixed generator

$$z_i = (b_1 x_{1,i} + \cdots + b_J x_{J,i}) \bmod M_1,$$

where $b_1, \ldots, b_J$ are integers and M_1 is the largest of all M_j. It appears that sequences generated by such a generator are equivalent to sequences generated by a single generalized congruential generator with M, which is the product of M_j, but with better statistical properties.

As mixed generators, we can also use the Fibonacci generators

$$x_i = x_{i-n} \odot x_{i-k} \bmod M \text{ for some } n \text{ and } k,$$

where $\odot$ denotes addition, subtraction, or multiplication.

To guarantee the "good" properties of such generators a long seed is required, which has to be generated by another generator.

Nowadays in common use is the generator by Matsumoto and Nishimura [39] called Mersenne Twister. This generator has period $(2^{19937}-1)$, and no correlation has been observed for vectors to dimension 623.

2.2 Non-uniform variates

Pseudo-random number generators usually produce samples from a uniform distribution. To obtain samples from other distributions, we have to apply certain transformations.

Discrete distribution

To obtain a sample from the discrete distribution X

$$\mathbb{P}(X = a_i) = p_i, \quad i = 1, \ldots, n,$$

we apply the following algorithm:

1. Compute $c_k = \sum_{i=1}^{k} p_i$, $k = 1, \ldots, n$;
2. Generate $u \sim \mathcal{U}(0,1)$;
3. Find the smallest k such that $u \leq c_k$;
4. Put $Z = a_k$.

Such Z is a sample from the distribution of X.

Inversion

THEOREM. 2.5 *Let a random variable X possess a continuous and strictly increasing cumulative distribution function F_X and $u \sim \mathcal{U}(0,1)$. Then $F_X^{-1}(u)$ is a sample of X.*

Proof. The condition $u \sim \mathcal{U}(0,1)$ means $\mathbb{P}_{\mathcal{U}}(u \leq \xi) = \xi$ for $\xi \in [0,1)$. Hence, $\mathbb{P}_{\mathcal{U}}(F_X^{-1}(u) \leq x) = \mathbb{P}_{\mathcal{U}}(u \leq F_X(x)) = F_X(x)$. ■

Acceptance–rejection method

Let X be a random variable with density $f(x)$, but sampling from that density be complicated (large computational complexity). Assume that there is another random variable Y with density $g(x)$ which is easily simulated. Let $f(x) \le C\, g(x)$ for a constant $C < +\infty$.

The following algorithm generates samples of X:

1. Generate x from the density of Y and $u \sim \mathcal{U}(0,1)$.

2. If $u \le \frac{f(x)}{C\, g(x)}$, accept x as a sample of X. Otherwise return to p. 1.

THEOREM. 2.6 *The described above algorithm generates a sample of X.*

Proof. Let

$$A = \Big\{\omega\colon \mathcal{U}(\omega) \le \frac{f(Y(\omega))}{C\, g(Y(\omega))}\Big\}.$$

Then

$$\begin{aligned}\mathbb{P}(Y \in dx|A) &= \frac{\mathbb{P}(A \cap \{Y \in dx\})}{\mathbb{P}(A)} = \frac{g(x)\frac{f(x)}{C\, g(x)}dx}{\int_{\mathbb{R}} g(y)\frac{f(y)}{C\, g(y)}dy} \\ &= \frac{f(x)dx}{\int_{\mathbb{R}} f(y)dy} = f(x)dx = \mathbb{P}(X \in dx),\end{aligned}$$

where the value of $\mathbb{P}(A)$ is obtained by the following computation

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}\Big(\mathcal{U} \le \frac{f(Y)}{C\, g(Y)}\Big) = \mathbb{E}\Big(\mathbb{P}\Big(\mathcal{U} \le \frac{f(Y)}{C\, g(Y)}\Big|Y\Big)\Big) \\ &= \mathbb{E}\Big(\frac{f(Y)}{C\, g(Y)}\Big) = \int_{\mathbb{R}} \frac{f(y)}{C\, g(y)} g(y)dy = \frac{1}{C}\int_{\mathbb{R}} f(y)dy.\end{aligned}$$

■

Transformation of random variables

The theorem below follows easily from the change of variables theorem.

THEOREM. 2.7 *Let X be a random variable in $\mathbb{R}^d$ with a positive density ϕ supported on S. Assume that a transformation g: $S \to B$, $S, B \subset \mathbb{R}^d$, is invertible*

and that the inverse is continuously differentiable on B (is $C^1(B)$). Then $Y = g(X)$ has the density

$$\phi\big(g^{-1}(y)\big)\left|\frac{\partial(x_1,\ldots,x_d)}{\partial(y_1,\ldots,y_d)}\right|,\ y\in B,$$

where $x = g^{-1}(y)$ and $\frac{\partial(x_1,\ldots,x_d)}{\partial(y_1,\ldots,y_d)}$ denotes the Jacobian matrix of g^{-1}.

We can apply the transformation method to generate samples from a normal distribution.

Example. 2.8 (Box-Muller algorithm) *We define the transformation*

$$\begin{aligned} y_1 &= \sqrt{-2\ln x_1}\cos 2\pi x_2 = g_1(x_1,x_2),\\ y_2 &= \sqrt{-2\ln x_1}\sin 2\pi x_2 = g_2(x_1,x_2), \end{aligned}$$

for $(x_1,x_2)\in[0,1]^2$.

The inverse transformation is given by

$$\begin{aligned} x_1 &= \exp\Big(-\frac{1}{2}(y_1^2+y_2^2)\Big),\\ x_2 &= \frac{1}{2\pi}\arctan\frac{y_2}{y_1}. \end{aligned}$$

The Jacobian matrix is as follows

$$\begin{aligned} \det\begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2}\\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2}\end{pmatrix} &= \frac{1}{2\pi}\exp\Big(-\frac{1}{2}(y_1^2+y_2^2)\Big)\left(-y_1\frac{1}{1+\frac{y_2^2}{y_1^2}}\frac{1}{y_1} - y_2\frac{1}{1+\frac{y_2^2}{y_1^2}}\frac{y_2}{y_1^2}\right)\\ &= -\frac{1}{2\pi}\exp\Big(-\frac{1}{2}(y_1^2+y_2^2)\Big). \end{aligned}$$

The above computations show that $\left|\det\left(\frac{\partial(x_1,x_2)}{\partial(y_1,y_2)}\right)\right|$ is the density of a 2-dimensional standardized normal variable.

When $X\sim\mathcal{U}(0,1)^2$, then $Y=(g_1(X),g_2(X))$ is a 2-dimensional standardized normal variable and Y_1, Y_2 are i.i.d.

Remark. 2.1 (Neave effect) *In 1973, H. R. Neave [43] discovered a surprising result of applying the Box-Muller algorithm to a sample obtained from a linear congruential generator. Since the density of normal distribution is supported on the whole $\mathbb{R}$, we can expect that the pairs (Y_1,Y_2) will cover the whole $\mathbb{R}^2$. Contrary to that expectation the generated pairs fall into a small range (rectangle)*

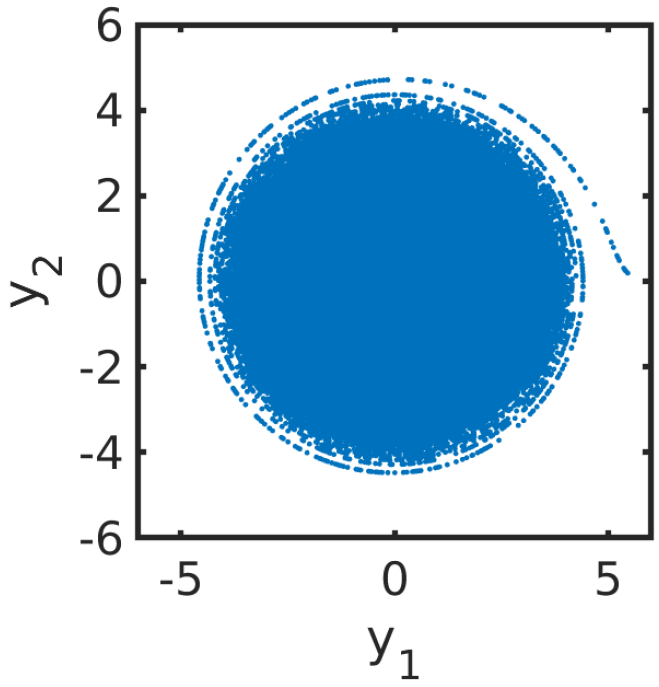

Figure 2.4: The Neave effect for the Box-Muller algorithm.

around zero. This surprising behavior is illustrated in Figure 2.4 obtained for 10^7 *simulations. The Neave effect has been observed for the Box-Muller algorithm and other uniform generators. There is a suspicion that the effect can also occur for other random number generators.*

Example. 2.9 (Marsaglia's polar method) *The Box-Muller algorithm has been improved by Marsaglia to avoid the use of trigonometric functions. That modification is important, since the computation of trigonometric functions is very time-consuming. Marsaglia's algorithm replaces the evaluation of trigonometric functions in the Box-Muller algorithm by the acceptance-rejection method:*

1. *Generate* $u_1, u_2 \sim \mathcal{U}(0,1)$.
2. *Perform transformation* $v_1 = 2u_1 - 1$, $v_2 = 2u_2 - 1$. *Then* $v_1, v_2 \sim \mathcal{U}(-1,1)$.
3. *Accept a pair* (v_1, v_2), *when* $v_1^2 + v_2^2 \leq 1$. *Such a pair is uniformly distributed on* $D = \{(v_1, v_2)\colon\ v_1^2 + v_2^2 \leq 1\}$.

4. *Every pair $(v_1, v_2) \in D$ can be considered as defined by the polar coordinates (z, θ):*

$$v_1 = z\cos\theta, \quad v_2 = z\sin\theta.$$

The variates $w = z^2$ and θ are independent: $w \sim \mathcal{U}(0,1)$, $\theta \sim \mathcal{U}(0, 2\pi)$.

5. *The variables*

$$y_1 = \sqrt{-\frac{\ln w}{w}}v_1, \quad y_2 = \sqrt{-\frac{\ln w}{w}}v_2$$

are normally distributed and independent.

The pair y_1, y_2 is normally distributed by the Box-Muller algorithm. Since $\theta = \arctan\frac{v_2}{v_1}$, then, taking $x_1 = v_1^2 + v_2^2 \equiv w$ and $x_2 = \frac{1}{2\pi}\theta$, we obtain $x_1, x_2 \sim \mathcal{U}(0,1)$, where $\cos(2\pi x_2) = \frac{v_1}{\sqrt{w}}$ and $\sin(2\pi x_2) = \frac{v_2}{\sqrt{w}}$. These transformations reduce the Marsaglia algorithm to the Box-Muller algorithm.

Example. 2.10 *The Box-Muller and Marsaglia algorithms are not sufficiently accurate for financial applications. There are many algorithms using inversion to provide more accurate normal distributions. As an example, we present the modification by Moro [41] of the Beasley-Springer algorithm [4] with accuracy 3×10^{-9}. But there are algorithms with higher accuracy, and compatible computational complexity (cf. [48]). For $0.5 \le y < 0.92$, the Beasley-Springer-Moro algorithm uses the formula*

$$F^{-1}(y) \approx \frac{\sum_{n=0}^{3} a_n (y-0.5)^{2n+1}}{1 + \sum_{n=0}^{3} b_n (y-0.5)^{2n+2}},$$

and for $y \ge 0.92$, the formula

$$F^{-1}(y) \approx \sum_{n=0}^{8} c_n \Big(\ln\big(-\ln(1-y)\big)\Big)^n.$$

Constants for the Beasley-Springer-Moro algorithm are given in Table 2.1. The case $0 \le y \le 0.5$ is handled by symmetry.

2.3 Multivariate random variables

Multivariate distributions are usually obtained from one-dimensional distributions. But only in very particular situations, multivariate distributions can be obtained as the Cartesian product of univariate distributions.

a0 =	2.50662823884	b0 =	-8.47351093090
a1 =	-18.61500062529	b1 =	23.08336743743
a2 =	41.39119773534	b2 =	-21.06224101826
a3 =	-25.44106049637	b3 =	3.13082909833
c0 =	0.3374754822726147	c5 =	0.0003951896511919
c1 =	0.9761690190917186	c6 =	0.0000321767881768
c2 =	0.1607979714918209	c7 =	0.0000002888167364
c3 =	0.0276438810333863	c8 =	0.0000003960315187
c4 =	0.0038405729373609		

Table 2.1: Constants for the Beasley-Springer-Moro algorithm.

Uniform distribution

A multidimensional uniform distribution can be obtain as a tuple of independent one-dimensional distributions $(U_1, \ldots, U_d)$, where $U_i \sim \mathcal{U}(0,1)$.

Normal variates

A standardized d-dimensional normal distribution $\mathcal{N}_d(0, I_d)$, where I_d is d-dimensional identity matrix, can be obtained from one-dimensional normal distributions

$$X \sim \mathcal{N}_d(0, I_d) \Longleftrightarrow X = (X_1, \ldots, X_d), \text{ where } X_i \sim \mathcal{N}(0,1),\ X_i \text{ are i.i.d.}$$

This is due to the form of the density of d-dimensional normal variable $\mathcal{N}_d(\mu, \Sigma)$

$$\phi(x) = \frac{1}{(2\pi)^{d/2}} \frac{1}{(\det \Sigma)^{1/2}} \exp\Big(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\Big).$$

For $\mu = 0$ and $\Sigma = I_d$ this density is the product of standardized one-dimensional densities. Knowing the density of a correlated normal distribution, we can design a transformation of a standardized normal distribution $\mathcal{N}_d(0, I_d)$, which can be used to generate samples from the correlated distribution.

LEMMA. 2.11 *Let $Z \sim \mathcal{N}_d(0, I_d)$ and A be a nonsingular matrix of dimension $d \times d$. Then $AZ \sim \mathcal{N}_d(0, AA^\top)$ and $\mu + AZ \sim \mathcal{N}_d(\mu, AA^\top)$.*

Proof. Let $X = AZ$ and $\phi(z)$ be the density of Z. By the substitution $x = Az$ we obtain

$$\exp\Big(-\frac{1}{2}z^\top z\Big) = \exp\Big(-\frac{1}{2}(A^{-1}x)^\top (A^{-1}x)\Big) = \exp\Big(-\frac{1}{2}x^\top (A^{-1})^\top A^{-1} x\Big).$$

Changing variables we get

$$\int_{\mathbb{R}^d} \phi(z)dz = \int_{\mathbb{R}^d} \phi(x)|\det A|^{-1}dx$$
$$= \int_{\mathbb{R}^d} \frac{1}{|\det A|}\frac{1}{(2\pi)^{d/2}} \exp\Big(-\frac{1}{2}x^\top \big(AA^\top\big)^{-1}x\Big).$$

Hence,

$$\frac{1}{|\det A|}\frac{1}{(2\pi)^{d/2}} \exp\Big(-\frac{1}{2}x^\top \big(AA^\top\big)^{-1}x\Big)$$

is the density of $\mathcal{N}_d(0, AA^\top)$. ■

To generate a sample from $\mathcal{N}_d(\mu, \Sigma)$ we have to find a matrix A, such that $\Sigma = AA^\top$. The covariance matrix Σ is symmetric and positive definite. Hence, we find A by the Cholesky decomposition of Σ.

Remark. 2.2 *To generate samples from $\mathcal{N}_d(\mu, \Sigma)$ we can use spectral decomposition of Σ. Since Σ is symmetric, positive definite matrix, it has d positive eigenvalues and d eigenvectors which span $\mathbb{R}^d$ and $\Sigma = \Gamma\Lambda\Gamma^\top$, where Λ is the diagonal matrix of eigenvalues and Γ is the matrix of eigenvectors. If $Z \sim \mathcal{N}_d(0, I_d)$, then $\mu + \Gamma\Lambda^{1/2}Z \sim \mathcal{N}_d(\mu, \Sigma)$. Let us observe that $\Lambda^{1/2}$ is well defined due to the positivity of eigenvalues.*

Remark. 2.3 *In general, the Cholesky decomposition and the spectral decomposition construction do not give same results. Indeed*

$$\Sigma = AA^\top = \Gamma\Lambda\Gamma^\top = \Gamma\Lambda^{1/2}\Lambda^{1/2}\Gamma^\top,$$

and

$$\Sigma(A^\top)^{-1} = A = \Gamma\Lambda^{1/2}\Lambda^{1/2}\Gamma^\top(A^\top)^{-1}$$

holds. But in general

$$\Lambda^{1/2}\Gamma^\top(A^\top)^{-1} \neq I_d.$$

Hence

$$AZ \neq \Gamma\Lambda^{1/2}Z$$

although both methods generate samples from $\mathcal{N}_d(\mu, \Sigma)$.

In implementations, the Cholesky decomposition is more effective. In the Cholesky factorization, the matrix A is lower triangular. It makes calculations of AZ particularly convenient because it reduces the calculations complexity by the factor of 2 compared to the multiplication of Z by the full matrix $\Gamma\Lambda^{1/2}$. In

addition, the error propagates much slower in the Cholesky factorization. There are, however, situations in which using spectral decomposition gives some advantages. The eigenvalues and eigenvectors of the covariance matrix have a statistical interpretation that is sometimes useful. Examples of such uses are in some variance reduction methods.

Other multi-dimensional distributions

Uniform or normal distributions are rather exceptional examples of designing multivariate distributions as products of one-dimensional distributions. In the majority of situations generating samples from multivariate distributions requires special algorithms.

A good example is an algorithm of sampling from d-dimensional Student t-distribution $t_d(\nu, \mu, \Sigma)$, where ν is the number of degrees of freedom, μ – location vector, and Σ – dispersion matrix. To generate samples from the d-dimensional Student t-distribution we can use the formula

$$t_d(\nu, \mu, \Sigma) \sim \frac{\mathcal{N}_d(\mu, \Sigma)}{\sqrt{\chi^2_\nu / \nu}},$$

where χ^2_ν denotes an independent chi-square distribution with ν degrees of freedom. Using the above relation to generate a sample from $t_d(\nu, \mu, \Sigma)$ we have to generate two independent samples: one from $\mathcal{N}_d(\mu, \Sigma)$ and another from χ^2_ν.

As another example, we can consider sampling from d-dimensional asymmetric Laplace distribution $AL_d(m, \Sigma)$. This class of distributions is suitable for modeling heavy tailed multivariate data, which retain the finiteness of moments. To generate samples from $Y \sim AL_d(m, \Sigma)$ we generate a sample from $\mathcal{N}_d(0, \Sigma)$, an independent sample from a standard exponential distribution $Ex(1)$, and use the representation of the asymmetric Laplace distribution

$$AL_d(m, \Sigma) \sim mEx(1) + \sqrt{Ex(1)}\mathcal{N}_d(0, \Sigma).$$

2.4 Low discrepancy sequences

Assume that we have to compute a definite integral $\int_0^1 f(x)dx$. When function $f(x)$ is complicated, we can approximate the value of this integral by a numerical procedure. To this end, we can use the Monte Carlo method, which gives the approximation error $O(N^{-1/2})$ (see Chapter 1). On the other hand, we can apply a deterministic quadrature dividing interval $[0, 1]$ into N subintervals of equal length and using trapezoidal approximation. The error of such an approximation is

$O(N^{-1})$, which is much better than the Monte Carlo error (in addition, this error is deterministic contrary to the Monte Carlo error, which is of statistical nature).

On the other hand, the error estimate obtained in advance gives only the order of error magnitude. The exact value can be reasonably approximated only after computations. If the computed error is larger than the assumed accuracy, we have to repeat computations with a larger N. For the deterministic trapezoidal method, that requires the repetition of the whole computations for a new larger N. The Monte Carlo method is much more flexible: we sample additional numbers and perform computations only for these additional numbers, preserving the previously computed values. The idea of low discrepancy sequences is a compromise between numerical quadratures giving deterministic error and advantages of Monte Carlo which enables free placing of sample points until the desired accuracy is met. The most important advantage of low discrepancy sequences is the improvement of error estimates to $O(N^{-(1-\epsilon)})$ independently of the problem dimension which is significantly better than in Monte Carlo methods (in fact, the constant in the error estimate depends on the dimension, and the picture is not as good as the expression $O(N^{-(1-\epsilon)})$ can suggest).

DEFINITION. 2.12 *Let $\{u_n\}_{n\geq 1}$ be a sequence of points in $[0,1]^d$. This point set is called* uniformly distributed *in $[0,1]^d$, if for each cube $Q_y = \{x \in [0,1]^d\colon\ 0 \leq x_i \leq y_i, i = 1,\ldots,d\}$ defined by $y \in [0,1]^d$, the equality holds*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \chi_{Q_y}(u_n) = vol(Q_y).$$

DEFINITION. 2.13 *For a sequence $u = \{u_1,\ldots,u_N\}$ of points from $[0,1]^d$ and Q_y, $y \in [0,1]^d$, we define*

$$F(y) = vol(Q_y), \quad F_N^u(y) = \frac{1}{N} \sum_{n=1}^{N} \chi_{Q_y}(u_n).$$

The star discrepancy *is given by*

$$D_N^*(u) = \sup_{y\in[0,1]^d} |F(y) - F_N^u(y)|.$$

Remark. 2.4 *By the definition of discrepancy if $u = \{u_1,\ldots,u_N\}$ is a subsequence of a uniformly distributed sequence $\{u_n\}_{n\geq 1}$, then $\lim_{N\to\infty} D_N^*(u) = 0$.*

THEOREM. 2.14 (Koksma-Hlawka inequality) *Let $f\colon\ [0,1]^d \to \mathbb{R}$ be a function of bounded variation. For each sequence $u = \{u_1,\ldots,u_N\}$ in $[0,1]^d$*

$$\left| \int_{[0,1]^d} f(x)dx - \frac{1}{N} \sum_{n=1}^{N} f(u_n) \right| \leq V(f) D_N^*(u),$$

where $V(f)$ is the Hardy-Krause variation of f in $[0,1]^d$.

Proof. For simplicity we will consider only a one-dimensional case assuming in addition that f is of class $C^1([0,1])$ (for a more complete proof see [16]). Then the Hardy-Krause variation is the total variation which for $f \in C^1([0,1])$ is equal $V(f) = \int_0^1 |f'(x)|dx$.

Let us take a sequence of points $u = \{u_1, \ldots, u_N\}$ in $[0,1]$. Using identity $f(x) = f(1) - \int_x^1 f'(y)dy$ we obtain

$$\begin{aligned}\int_0^1 f(x)dx - \frac{1}{N}\sum_{n=1}^N f(u_n) &= \frac{1}{N}\sum_{n=1}^N \int_{u_n}^1 f'(y)dy - \int_0^1 \int_x^1 f'(y)dy\, dx \\ &= \int_0^1 \frac{1}{N}\sum_{n=1}^N \mathbb{1}_{(u_n,1]}(y) f'(y)dy - \int_0^1 \int_0^y f'(y)dx\, dy \\ &= \int_0^1 f'(y)\left(\frac{1}{N}\sum_{n=1}^N \mathbb{1}_{(u_n,1]}(y) - y\right)dy.\end{aligned}$$

Let us observe that for a given y

$$\frac{1}{N}\sum_{n=1}^N \mathbb{1}_{(u_n,1]}(y) = \frac{1}{N}\sum_{n=1}^N \mathbb{1}_{[0,y)}(u_n).$$

The right hand side of this equality is $F_N^u(y)$ by Definition 2.13. Since in one dimension $F(y) = y$, we obtain

$$\int_0^1 f(x)dx - \frac{1}{N}\sum_{n=1}^N f(u_n) = \int_0^1 f'(y)\big(F_N^u(y) - F(y)\big)dy.$$

Hence

$$\begin{aligned}\left|\int_0^1 f(x)dx - \frac{1}{N}\sum_{n=1}^N f(u_n)\right| &\le \int_0^1 |f'(y)|\big|(F_N^u(y) - F(y)\big|dy \\ &\le \sup_{y\in[0,1]}\big|(F_N^u(y) - F(y)\big| \int_0^1 |f'(y)|dy = V(f)D_N^*(u).\end{aligned}$$

■

DEFINITION. 2.15 *A sequence $u = \{u_1, \ldots, u_N\}$ in $[0,1]^d$ is called the* low discrepancy sequence*, if*

$$D_N^*(u) = O\left(\frac{(\ln N)^d}{N}\right).$$

Remark. 2.5 *One can prove that if $U = \{U_1, \dots, U_N\}$ is a sequence of independent random variables with uniform distribution on $[0,1]^d$, then samples from this sequence are uniformly distributed on $[0,1]^d$. For the discrepancy of such samples, we have the estimate*

$$\mathbb{E}\big(D_N^*(U)\big) = \frac{C_d}{\sqrt{N}}, \text{ where } C_d = O(\ln \ln N),$$

which shows that the discrepancy of a sequence of random variables is of order $N^{-1/2}$.

Example. 2.16 (Halton sequences) *This is an example of an easily computed sequence of low discrepancy.*

Let $b \geq 2$ be a prime number. As is well known, each integer n can be expanded in the basis b

$$n = a_0 + a_1 b + \cdots + a_k b^k, \quad \text{where } a_i \in \{0, 1, \dots, b-1\}.$$

For n with the above expansion we define the radical-inverse function

$$\Psi_b(n) = \frac{a_0}{b} + \frac{a_1}{b^2} + \cdots + \frac{a_k}{b^{k+1}}.$$

Then the b-adic van der Corput sequence *is defined as the one-dimensional sequence $\{\Psi_b(n)\}_{n\in\mathbb{N}}$.*

Van der Corput sequences are used as ingredients in the construction of Halton sequences. The Halton sequence *of dimension d is a sequence $u = \{u_i\}$ in $[0,1]^d$ defined for $b_1, \dots, b_d$ pairwise prime numbers as the sequence of vectors*

$$u_i = \big(\Psi_{b_1}(i), \Psi_{b_2}(i), \dots, \Psi_{b_d}(i)\big).$$

One can prove (but the proof is not easy, see [16]) that the Halton sequence is a sequence of low discrepancy

$$D_N^*(u) \leq C\frac{(\ln N)^d}{N}, \text{ for a constant } C > 0.$$

Although the Halton sequence is a low discrepancy sequence, it is not advised for computation of high dimensional integrals since in pairs of the van der Corput sequences for large bases occur cycles with decreasing periods. The Halton sequence is a good choice for moderate dimensions where the bases of the van der Corput sequences are not too large. For higher dimensions, one can use the Sobol', Faure, or Niederreiter sequences which are obtained by permuting terms

in the van der Corput sequences for small bases (in the construction of Sobol' sequences only number representation in basis 2 is used). The construction of these sequences is rather involved and its detailed description is beyond the scope of these lecture notes. The interested reader can consult the book by Glasserman [21] and the references cited herein.

The computational method, which uses as nodes of approximation low discrepancy sequences, is called the *Quasi Monte Carlo* method. From the Koksma-Hlawka inequality and the definition of low discrepancy sequences, we obtain an error estimate for this method $O(N^{-(1-\epsilon)})$. Hence, the estimate is better than for the Monte Carlo method with nodes obtained from pseudo-random numbers. Since by Quasi Monte Carlo methods, we can compute only multidimensional integrals, then we need to transform financial problems into such integrals.

Chapter 3

Monte Carlo methods

The history of Monte Carlo methods goes back to Stanislaw Ulam who working on the Manhattan Project suggested to John von Neumann that the newly developed ENIAC computer would give them the means to carry out calculations based on statistical sampling.

The name "the Monte Carlo method" is attributed to their coworker Nicholas Metropolis who was partly inspired by Ulam's anecdotes of his gambling uncle who "just had to go to Monte Carlo". In print the name has appeared for the first time in the paper: N. Metropolis, S. Ulam – The Monte Carlo method, *Journal of American Statistical Association*, 44 (1949). Phelim Boyle [8] was the first who in 1977 used Monte Carlo methods in quantitative finance to compute option prices.

3.1 Monte Carlo integration

Consider the Monte Carlo (MC) computation of expected value $\mathbb{E}(X)$ of a random variable X with a known distribution. The simplest algorithm, the so-called *crude Monte Carlo*, can be summarized as follows:

1. Generate a sample $X_1, \ldots, X_N$ from the distribution of X.

2. Compute the sample average

$$\hat{X} = \frac{1}{N} \sum_{i=1}^{N} X_i.$$

The following theorem, which follows from the strong law of large numbers, gives a theoretical foundation for MC simulations.

THEOREM. 3.1 *Let $X_1, \ldots, X_N$ be an i.i.d. sample from the distribution of X, where $\mathbb{E}(X) = \mu$, $\mathrm{Cov}(X) = \Sigma$. Let $\hat{X}_N$ denote the sample mean value and $\hat{\Sigma}_N$, the sample covariance matrix (subscript N indicates that these moments are computed from an N-element sample). Then:*

1. $\mathbb{E}(\hat{X}_N) = \mu$.
2. $\mathbb{E}(\hat{\Sigma}_N) = \Sigma$.
3. $\lim_{N\to\infty} \hat{X}_N = \mu$, *a.s.*
4. $\lim_{N\to\infty} \hat{\Sigma}_N = \Sigma$, *a.s.*

By the central limit theorem, we have the following corollary.

COROLLARY. 3.2 *If X is one-dimensional ($\Sigma = \sigma^2$), then*

$$\sqrt{N}\big(\hat{X}_N - \mathbb{E}(X)\big) \to \mathcal{N}(0, \sigma^2) \text{ in distribution.}$$

Therefore, if z_α denotes α-quantile of a standardized normal distribution, then

$$\frac{\sigma z_{\alpha/2}}{\sqrt{N}} < \hat{X}_N - \mathbb{E}(X) < \frac{\sigma z_{1-\alpha/2}}{\sqrt{N}}.$$

DEFINITION. 3.3 *The confidence interval with a confidence level α for $\hat{X}_N$ is*

$$\Big(\hat{X}_N - \frac{\sigma z_{1-\alpha/2}}{\sqrt{N}}, \hat{X}_N - \frac{\sigma z_{\alpha/2}}{\sqrt{N}}\Big).$$

Since $z_{1-\alpha/2} = -z_{\alpha/2}$, this confidence interval can be written as

$$\Big(\hat{X}_N - \frac{\sigma z_{1-\alpha/2}}{\sqrt{N}}, \hat{X}_N + \frac{\sigma z_{1-\alpha/2}}{\sqrt{N}}\Big).$$

Remark. 3.1 *In real simulations, we do not know the true value of variance σ and we replace that value with the sample variance*

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{j=1}^{N}(X_j - \hat{X})^2 = \frac{1}{N-1}\sum_{i=j}^{N} X_j^2 - \frac{N}{N-1}\hat{X}^2.$$

Let Φ be the cumulative distribution function of a standardized normal variable. Then we have the identity

$$\Phi(z_{1-\alpha/2}) - \Phi(z_{\alpha/2}) = (1 - \alpha/2) - \alpha/2 = 1 - \alpha.$$

Hence, if I_α is the confidence interval with a confidence level α, then $\mathbb{P}(\mathbb{E}(X) \in I_\alpha) = 1 - \alpha$, i.e., the confidence interval I_α contains the true value of $\mathbb{E}(X)$ with probability $(1 - \alpha)$.

MC simulations of $\mathbb{E}$(g(X))

W have earlier remarked that for a random variable X with a known density ϕ_X the computation of $\mathbb{E}(g(X))$ can be reduced to

$$\mathbb{E}\big(g(X)\big) = \int_{\mathbb{R}^d} g(x)\phi_X(x)dx = \int_{[0,1]^d} f(x)dx,$$

which can be treated as the mean value $\mathbb{E}(f(Y))$ for a random variable Y with uniform distribution on $[0,1]^d$. The simulation algorithm of crude Monte Carlo is then:

1. Generate a sample $Y_1, \ldots, Y_N$ from the uniform distribution of Y on $[0,1]^d$.
2. Compute the average $\hat{f} = \frac{1}{N}\sum_{j=1}^{N} f(Y_j)$, which approximates the integral.

THEOREM. 3.4 *Let $f \in L^2([0,1]^d)$ and*

$$V^2(f) = \int_{[0,1]^d} f^2(x)dx - \left(\int_{[0,1]^d} f(x)dx\right)^2 < \infty.$$

$\hat{f}$ computed by the crude Monte Carlo algorithm has the following properties:

1. *$\hat{f} \to \mathbb{E}(f(Y))$ for $N \to \infty$, a.s.*
2. *Let $\hat{\delta} = \int_{[0,1]^d} f(x)dx - \hat{f}$ then $\mathrm{Var}(\hat{\delta}) = \frac{V^2(f)}{N}$.*

Proof. If $Y_1, \ldots, Y_N$ are i.i.d., then $f(Y_1), \ldots, f(Y_N)$ are i.i.d. too, since $f \in L^2$ can be approximated by simple functions. The convergence $\hat{f} \to \mathbb{E}(f(Y))$ follows from the strong law of large numbers.

To compute $\mathrm{Var}(\hat{\delta})$ we observe that $\mathbb{E}(\hat{\delta}) = 0$ which follows from the equality $\mathbb{E}(\hat{f}) = \mathbb{E}(f(Y))$ valid by the i.i.d. property of $Y_1, \ldots, Y_N$. Then we get

$$\mathrm{Var}(\hat{\delta}) = \mathbb{E}(\hat{\delta}^2) = \mathbb{E}\left(\int_{[0,1]^d} f(x)dx - \frac{1}{N}\sum_{j=1}^{N} f(Y_j)\right)^2.$$

Let us denote $v(x) := \int_{[0,1]^d} f(x)dx - f(x)$. Then

$$\mathrm{Var}(\hat{\delta}) = \mathbb{E}\left(\frac{1}{N^2}\Big(\sum_{j=1}^{N} v(Y_j)\Big)^2\right),$$

due to the equality $\int_{[0,1]^d} f(x)dx = \frac{1}{N}\sum_{j=1}^{N}\int_{[0,1]^d} f(x)dx$.

Since $v(Y_j)$ are i.i.d., we have $\mathbb{E}\big(v(Y_i)v(Y_j)\big) = 0$, for $i \neq j$. These equalities supplemented with $\mathbb{E}(v(Y_j)) = 0$ give

$$\mathrm{Var}(\hat{\delta}) = \frac{1}{N^2}\sum_{j=1}^{N}\mathbb{E}\big(v^2(Y_j)\big) = \frac{1}{N}\mathrm{Var}\big(v(Y)\big) = \frac{1}{N}\int_{[0,1]^d} v^2(x)dx.$$

On the other hand $\mathrm{Var}\big(v(Y)\big) = V^2(f)$, as

$$\mathrm{Var}\big(v(Y)\big) = \int_{[0,1]^d} v^2(x)dx = \int_{[0,1]^d} f^2(x)dx - \left(\int_{[0,1]^d} f(x)dx\right)^2.$$

■

3.2 Variance reduction methods

We have seen that the size of the confidence interval is determined by the value of $\sqrt{\mathrm{Var}(X)}/\sqrt{N}$. We would like to find a method to decrease the size of this interval by other means than increasing the sample size N, which is usually very costly (computational complexity is, in the majority of cases, a linear function of N). These methods are called variance reduction methods, and we describe a number of them. We illustrate these methods by examples taken from quantitative finance.

Importance sampling

Suppose that we want to compute $x = \mathbb{E}(X)$. The concept of importance sampling is to modify the distribution of X so that most of the sampling is done on those regions which contribute the most to x. We modify the initial distribution $\mathbb{P}(d\omega)$ of X to an importance sampling distribution $\tilde{\mathbb{P}}(d\omega)$ such that $x = \mathbb{E}(X) = \tilde{\mathbb{E}}(LX)$, where $L = \frac{d\mathbb{P}}{d\tilde{\mathbb{P}}}$ is the Radon-Nikodym derivative. The problem is to make an efficient choice of the modified distribution $\tilde{\mathbb{P}}(d\omega)$.

THEOREM. 3.5 *Let X be a random variable with a distribution $\mathbb{P}$. Let us define $\mathbb{P}^*$ by the Radon-Nikodym derivative*

$$\frac{d\mathbb{P}}{d\mathbb{P}^*} = \frac{\mathbb{E}|X|}{|X|} = L^*, \quad \textit{i.e.,} \quad \mathbb{P}^*(d\omega) = \frac{1}{L^*}\mathbb{P}(d\omega).$$

Then the importance sampling estimator of $x = \mathbb{E}(X)$ under measure $\mathbb{P}^*$ has smaller variance than any other estimator obtained by a change of measure. If $X \geq 0$, $\mathbb{P}$ $a.s.$, then the $\mathbb{P}^*$-variance of the importance sampling estimator is equal to 0.

Proof. Let $X_1, \ldots, X_N$ be an i.i.d. sample of X. The importance sampling estimator under measure $\mathbb{P}^*$ is

$$\hat{x}^* = \frac{1}{N}\sum_{j=1}^{N} X_j L^*(X_j).$$

Its variance is

$$\mathrm{Var}^*(\hat{x}^*) = \mathrm{Var}^*(XL^*).$$

We have by the definition of L^*

$$\mathbb{E}^*\Big(\big(XL^*\big)^2\Big) = \mathbb{E}^*\Big(|X|^2\big(L^*\big)^2\Big) = \mathbb{E}^*\Big(\big(\mathbb{E}|X|\big)^2\Big) = \big(\mathbb{E}|X|\big)^2.$$

If $\tilde{\mathbb{P}}$ is another measure with the Radon-Nikodym derivative $\tilde{L}$ such that $\mathbb{E}(X) = \tilde{\mathbb{E}}(X\tilde{L})$, then

$$\big(\mathbb{E}|X|\big)^2 = \Big(\tilde{\mathbb{E}}\big(|X|\tilde{L}\big)\Big)^2 \le \tilde{\mathbb{E}}\Big(\big(X\tilde{L}\big)^2\Big).$$

Hence

$$\mathbb{E}^*\Big(\big(XL^*\big)^2\Big) \le \tilde{\mathbb{E}}\Big(\big(X\tilde{L}\big)^2\Big).$$

Since $x = \mathbb{E}(X) = \mathbb{E}^*(XL^*) = \tilde{\mathbb{E}}(X\tilde{L})$, we obtain $\mathrm{Var}^*(XL^*) \le \widetilde{\mathrm{Var}}(X\tilde{L})$.

If $X \ge 0$, then

$$\mathrm{Var}^*(XL^*) = \mathbb{E}^*\big((XL^*)^2\big) - \Big(\mathbb{E}^*(XL^*)\Big)^2 = \big(\mathbb{E}(X)\big)^2 - \big(\mathbb{E}(X)\big)^2 = 0.$$

■

Remark. 3.2 *The choice of measure suggested by Theorem 3.5 can never be implemented in practice since to obtain L^* we have to know $\mathbb{E}|X|$ and this is the value, we want to estimate. Nevertheless, Theorem 3.5 suggests that variance reduction can be achieved by sampling in proportion to $|X(\omega)|$.*

Example. 3.6 *Suppose that we want to compute $z = \mathbb{P}(X \in A)$, where X is a d-dimensional Gaussian random vector with mean 0 and covariance matrix D. Theorem 3.5 suggests sampling in proportion to $\phi_X(x)\mathbb{1}_{x\in A}$, where ϕ_X is the density of X. The rapid decay of $\phi_X(x)$ for large x suggests sampling in the vicinity of point x^*, where ϕ_X has a maximum over A. Hence, the distribution $\tilde{\mathbb{P}}(d\omega)$ can be obtained by a shift of $\mathbb{P}(d\omega)$, with mean 0, to $\tilde{\mathbb{P}}(d\omega)$, with mean x^*. This gives $L = \frac{\phi_X}{\tilde{\phi}_X}$, where $\tilde{\phi}_X$ is the density with mean x^*. Knowing the Gaussian distribution formula we obtain*

$$L = \exp\Big(-(x^*)^\top D^{-1}X + \frac{1}{2}(x^*)^\top D^{-1}x^*\Big).$$

Example. 3.7 *A special case of the previous example is the computation of* VaR *of an investment portfolio. Let us recall*

$$\mathrm{VaR}_\alpha(X) = \inf_x \mathbb{P}(-X \leq x) \geq 1 - \alpha.$$

Usually, VaR *is estimated by computing* $\mathbb{P}(-X > x)$ *for a sequence of different* x *and choosing* x *corresponding to the prescribed confidence level* α*.*

For an investment portfolio of d *assets with prices* $S = (S_1, \ldots, S_d)$ *following a normal distribution, we have to compute* $\mathbb{P}(-X > x) = \mathbb{P}(-w^\top S > x)$*, where* w *are the weights of assets in the portfolio.*

Assume, like in the previous example, $S \sim \mathcal{N}(0, D)$*. To implement the importance sampling technique, we have to know* $x^* \in \mathbb{R}^d$*, a point in which the distribution of* S *has the maximum over* $-w^\top S > x$*. That leads to the following optimization problem*

$$\begin{cases} \max\limits_z \exp\Big(-\frac{1}{2} z^\top D^{-1} z\Big), \\ -w^\top z > x. \end{cases}$$

This problem is equivalent to

$$\begin{cases} \max\limits_z -\frac{1}{2} z^\top D^{-1} z, \\ -w^\top z > x, \end{cases}$$

which can be solved by the Lagrange multipliers giving

$$z^* = -\frac{x}{w^\top D w} D w.$$

Inserting that z^* *as* x^* *into the formula from the previous example we obtain the importance sampling change of measure in the MC simulations for* $\mathrm{VaR}_\alpha(X)$*.*

Remark. 3.3 *In practice we are interested in computing* VaR *of losses over a certain time interval. Let the loss be given by* $-X = V(t_0 + \Delta t) - V(t_0) = \Delta V$*, where* V *is the portfolio value. By a first-order approximation we get* $\Delta V \simeq \frac{\partial V}{\partial t}\Delta t + \frac{\partial V}{\partial S}\Delta S$*. Assuming that* ΔS *has a multivariate normal distribution, we obtain*

$$\mathbb{P}(-X > x) \simeq \mathbb{P}\Big(\frac{\partial V}{\partial S}\Delta S > x - \frac{\partial V}{\partial t}\Delta t\Big).$$

Hence, the problem is reduced to the problem analyzed in the last example.

Antithetic variates

To implement antithetic variates, we generate N i.i.d. random pairs

$$(X_1, X_2), (X_3, X_4), \ldots, (X_{2N-1}, X_{2N}).$$

We assume that variables X_{2j-1}, X_{2j} have identical variances and negative correlations, for $j = 1, \ldots, N$.

Then

$$\begin{aligned}\mathrm{Var}\Big(\frac{1}{2N}\sum_{i=1}^{2N} X_i\Big) =& \frac{1}{N}\mathrm{Var}\Big(\frac{X_1+X_2}{2}\Big) = \frac{1}{4N}\big(2\sigma^2(X_1) + 2\sigma^2(X_1)\rho\big)\\ =& \frac{1}{2N}\sigma^2(X_1)(1+\rho),\end{aligned}$$

where ρ is the correlation of X_1 and X_2.

Hence, we have to choose X_{2j-1}, X_{2j} to make ρ possibly close to -1. There is no simple method for such a choice. The following theorem is a step further to identifying a suitable candidate.

THEOREM. 3.8 *Let X be a one-dimensional random variable with a symmetric density $\phi(x)$, i.e., X and $-X$ have the same density. If g is a monotonic function and $\mathbb{E}\big(g(X)\big) < +\infty$, then*

$$\mathrm{Corr}\big(g(X), g(-X)\big) \le 0.$$

The above inequality is sharp if g is strictly monotonic over a set of positive $\mathbb{P}$-measure, where $\mathbb{P}$ is implied by the distribution of X.

Proof. Let $m = \mathbb{E}\big(g(X)\big) < +\infty$, $c = \inf\{y \in \mathbb{R}\colon\ g(y) \ge m\}$. Then we get

$$\begin{aligned}&\int_{\mathbb{R}} \big(g(x)-m\big)\big(g(-x)-m\big)\phi(x)dx\\ &\quad= \int_{\mathbb{R}} \big(g(x)-m\big)\big(g(-x)-g(-c)\big)\phi(x)dx\\ &\qquad+ \big(g(-c)-m\big)\int_{\mathbb{R}} \big(g(x)-m\big)\phi(x)dx.\end{aligned}$$

Since

$$\big(g(x)-m\big)\big(g(-x)-g(-c)\big) \le 0$$

and by definition $\int_{\mathbb{R}} \big(g(x)-m\big)\phi(x)dx = 0$, then

$$\mathrm{Cov}\big(g(X), g(-X)\big) = \int_{\mathbb{R}} \big(g(x)-m\big)\big(g(-x)-m\big)\phi(x)dx \le 0.$$

If $g(x)$ is strictly monotonic over $\phi(x) > 0$, then

$$\big(g(x) - m\big)\big(g(-x) - g(-c)\big) < 0, \text{ a.s.}$$

and the integral is strictly less than zero. ■

Example. 3.9 *A standard example of antithetic variates appears in the computation of option prices in the Black-Scholes model. Since the price of the underlying is*

$$X_T = X_0 \exp\Big(\big(r - \sigma^2/2\big)T + \sigma W_T\Big),$$

we take as an antithetic variate

$$X_T^- = X_0 \exp\Big(\big(r - \sigma^2/2\big)T - \sigma W_T\Big),$$

hence W_T is replaced by $-W_T$.

As W_T has symmetric density, for a monotonic $g(X_T)$ the function $f(W_T) = g(X_T)$ is monotonic as the superposition of a monotonic function g with the strictly increasing exponential function. Hence, for a monotonic g, the use of antithetic variates in the Black-Scholes model gives always a reduction of variance.

Control variates

The idea behind the control variates is as follows. We wish to estimate $x = \mathbb{E}(X)$. Suppose that we can somehow find another random variable Y, which is close to X in some sense and has known expectation $y = \mathbb{E}(Y)$.

Let $\hat{x}$ and $\hat{y}$ be estimators of x and y, respectively. Then the control variate estimator is

$$\hat{x}_{cv} = \hat{x} + \alpha(\hat{y} - y).$$

The optimal choice of α should minimize the variance of the control variate estimator

$$\mathrm{Var}(\hat{x}_{cv}) = \mathrm{Var}(\hat{x}) + \alpha^2 \mathrm{Var}(\hat{y}) + 2\alpha \mathrm{Cov}(\hat{x}, \hat{y}).$$

The right hand side is minimal for

$$\alpha = -\frac{\mathrm{Cov}(\hat{x}, \hat{y})}{\mathrm{Var}(\hat{y})}.$$

Then

$$\mathrm{Var}(\hat{x}_{cv}) = \mathrm{Var}(\hat{x}) - \frac{\big(\mathrm{Cov}(\hat{x}, \hat{y})\big)^2}{\mathrm{Var}(\hat{y})} = \mathrm{Var}(\hat{x})(1 - \rho^2),$$

where ρ is the correlation of $\hat{x}$ and $\hat{y}$.

This computation shows that we have to look for Y with $|\rho|$ possibly close to 1.

Example. 3.10 (Underlying asset as a control variate) *Suppose we are pricing an option in the Black-Scholes model. If S_t is an asset price, then $\tilde{S}_t = \exp(-rt)S_t$ is a martingale. Assume we are pricing a European option with payoff $Y = g(S_T)$ and maturity T. To estimate S_T we perform N simulations S^j, $j = 1, \dots, N$, and compute $Y^j = g(S^j)$. The control variate estimator is*

$$\frac{1}{N}\sum_{j=1}^{N}\Big(Y^j - \big(S^j - \exp(rT)S_0\big)\Big).$$

Example. 3.11 (Hedge control variates) *Because the payoff of a hedged portfolio has a lower standard deviation than the payoff of an unhedged one, using hedges can reduce the volatility of the portfolio. Let $V(t) = g(S_t)$ denote an option price at time t. A delta hedge consists of holding $\Delta = \partial V/\partial S$ shares of the underlying asset, which is rebalanced at discrete time intervals. At time T, the hedge consists of the savings account and the asset, which closely replicates the payoff of the option. If $\tilde{V}(t)$ is the discounted price then*

$$\tilde{V}(T) = \tilde{V}(t_0) + \int_{t_0}^{T} \frac{\partial \tilde{V}}{\partial \tilde{S}} d\tilde{S}.$$

We write a discrete approximation to the above formula dividing $[t_0, T]$ into n subintervals with endpoints t_i, $i = 0, \dots, n$, and replacing the integral by a discrete sum. In addition, we discount all terms to time T. Then we can write

$$V(T) = V(t_0)\mathrm{e}^{r(T-t_0)} + \sum_{i=0}^{n-1} \frac{\partial V(t_i)}{\partial S}\Big(S_{t_{i+1}} - S_{t_i}\mathrm{e}^{r(t_{i+1}-t_i)}\Big)\mathrm{e}^{r(T-t_{i+1})}.$$

Let us note that

$$\mathrm{CV} = \sum_{i=0}^{n-1} \frac{\partial V(t_i)}{\partial S}\Big(S_{t_{i+1}} - S_{t_i}\mathrm{e}^{r(t_{i+1}-t_i)}\Big)\mathrm{e}^{r(T-t_{i+1})}$$

is an approximation of a martingale with expectation zero (a stochastic integral). We will use this expression as a control variate. We simulate N trajectories $S^j_{t_i}$, $j = 1, \dots, N$. On each trajectory we compute the option price $V^j(T) = g\big(S^j_{t_n}\big)$ and the control variate

$$\mathrm{CV}^j = \sum_{i=0}^{n-1} \frac{\partial V^j(t_i)}{\partial S}\Big(S^j_{t_{i+1}} - S^j_{t_i}\mathrm{e}^{r(t_{i+1}-t_i)}\Big)\mathrm{e}^{r(T-t_{i+1})}.$$

Then we obtain the control variate estimator

$$V(t_0)\mathrm{e}^{r(T-t_0)} = \frac{1}{N}\sum_{j=1}^{N}\big(V^j(T) - \mathrm{CV}^j\big).$$

Example. 3.12 (Asian option in the Black-Scholes model) *Asian options are options with payoff depending on the average price of the underlying asset on* $[0,T]$

$$S_{ar} = \frac{1}{T}\int_0^T S_t dt.$$

Even in the most elementary Black-Scholes model, there is no analytic expression for the price of a call or put option on that underlying. On the other hand, for the corresponding geometric average Asian option

$$S_{gm} = \exp\Big(\frac{1}{T}\int_0^T \ln S_t dt\Big)$$

there is an analytic formula similar to the Black-Scholes formula.

To derive that formula we use the expression $S_t = S_0 \exp\Big((r-\sigma/2)t+\sigma W_t\Big)$. *Taking in that expression* $S_0 = 1$, *computing* $\ln S_t$ *and integrating, we obtain*

$$\frac{1}{T}\int_0^T \ln S_t\, dt = (r-\sigma^2/2)\frac{T}{2} + \frac{\sigma}{T}\int_0^T W_t dt.$$

To evaluate the integral $\int_0^T W_t dt$ *we observe that this random variable is normally distributed. It is sufficient to compute the mean and variance for* $X := \frac{\sigma}{T}\int_0^T W_t dt$. *It is easy to see that* $\mathbb{E}(X) = 0$ *due to* $\mathbb{E}(W_t) = 0$. *To evaluate the variance of* X *we compute only*

$$\mathbb{E}\big(X^2\big) = \frac{\sigma^2}{T^2}\mathbb{E}\Big(\int_0^T W_t dt\Big)^2.$$

Integrating by parts we obtain $\int_0^T W_t dt = \int_0^T (T-t) dW_t$. *Then*

$$\mathbb{E}\Big(\int_0^T W_t dt\Big)^2 = \mathbb{E}\Big(\int_0^T (T-t) dW_t\Big)^2 = \int_0^T (T-t)^2 dt = \frac{1}{3}T^3.$$

This gives the distribution of $\sigma/T\int_0^T W_t dt \sim \mathcal{N}(0, \sigma^2 T/3)$ *and also*

$$\frac{1}{T}\int_0^T \ln S_t dt \sim \mathcal{N}\Big(\big(r-\frac{\sigma^2}{2}\big)\frac{T}{2}, \frac{\sigma^2}{3}T\Big). \tag{3.1}$$

Hence, we can price the option using the Black-Scholes formula.

Let us recall that in the Black-Scholes model for a stock paying a continuous dividend with rate r_d, *we have*

$$\ln S_t \sim \mathcal{N}\Big(\big(r-r_d-\frac{v^2}{2}\big)t, v^2 t\Big), \tag{3.2}$$

where v is the stock volatility.

The call price in this model is given by the formula

$$V_t = \mathrm{e}^{-r_d(T-t)} S_t \Phi\big(d_1(S_t, T-t)\big) - K\mathrm{e}^{-r(T-t)} \Phi\big(d_2(S_t, T-t)\big), \qquad (3.3)$$

where

$$d_1(S_t, T-t) = \frac{\ln(S_t/K) + (r - r_d + \frac{v^2}{2})(T-t)}{v\sqrt{T-t}}$$
$$d_2(S_t, T-t) = d_1(S_t, T-t) - v\sqrt{T-t}.$$

Comparing the distribution given by (3.1) *with the distribution of* $\ln S_T$ *from* (3.2) *we see that by making in the Black-Scholes formula* (3.3) *the substitutions*

$$v = \frac{\sigma}{\sqrt{3}}, \quad r_d = \frac{r}{2} + \frac{\sigma^2}{12}, \quad r - r_d - \frac{1}{2}v^2 = \frac{r}{2} - \frac{\sigma^2}{4},$$

we arrive at the price of the geometric average Asian call option

$$V_0 = \mathrm{e}^{-\frac{1}{2}\left(r + \frac{\sigma^2}{6}\right)T} S_0 \Phi(b_1) - \mathrm{e}^{-rT} K \Phi(b_2),$$

where

$$b_1 = \frac{\ln \frac{S_0}{K} + \frac{1}{2}\left(r + \frac{\sigma^2}{6}\right)T}{\frac{\sigma}{\sqrt{3}}\sqrt{T}}, \quad b_2 = b_1 - \frac{\sigma}{\sqrt{3}}\sqrt{T}.$$

The price for the arithmetic average Asian option is estimated by the following algorithm: we discretize the time interval $[0, T]$ *with points* t_i*,* $i = 1, \ldots, n$*, simulate* N *price trajectories* $(S^j_{t_i})_{1 \le i \le n}$*,* $j = 1 \ldots, N$*, and on each trajectory compute*

$$A^j = \exp(-rT) \left(\frac{1}{n} \sum_{i=1}^{n} S^j_{t_i} - K \right)^+,$$
$$G^j = \exp(-rT) \left(\left(\prod_{i=1}^{n} S^j_{t_i} \right)^{\frac{1}{n}} - K \right)^+,$$
$$X^j = A^j - (G^j - V_0).$$

The estimator of the price for the arithmetic average Asian option is

$$\hat{X} = \frac{1}{N} \sum_{j=1}^{N} X^j.$$

Remark. 3.4 *The estimator of the last example is biased. This bias can be explained by the fact that in the MC simulation, we use discrete geometric averaging, and the analytic price has been computed for the continuous geometric average. Fortunately, we can derive (computations are a bit more complicated) an analytic formula for the discrete geometric average call option. Discretizing the time interval $[0,T]$ into n subintervals of equal length we obtain the following price of the discrete geometric average Asian call option*

$$V_D = \mathrm{e}^{-rT+\frac{n+1}{2n}(r-\sigma^2/2)T+\frac{(n+1)(2n+1)}{12n^2}\sigma^2 T} S_0\Phi(\hat{b}_1) - \mathrm{e}^{-rT}K\Phi(\hat{b}_2),$$

where

$$\hat{b}_1 = \frac{\log\frac{S_0}{K} + \frac{n+1}{2n}(r-\sigma^2/2)T + \frac{(n+1)(2n+1)}{6n^2}\sigma^2 T}{\sigma\sqrt{\frac{T(n+1)(2n+1)}{6n^2}}},$$

$$\hat{b}_2 = \hat{b}_1 - \sigma\sqrt{\frac{T(n+1)(2n+1)}{6n^2}}.$$

Using V_D instead of V_0 we remove the bias.

3.3 Greeks

In this section, we will describe selected methods for estimating sensitivities, i.e., derivatives with respect to parameters for expectations of certain random variables. When these expectations are contingent claim prices, the sensitivities are called Greeks.

Finite differences

Consider a contingent claim $Y(\theta)$ depending on a parameter θ. Our goal is to compute the derivative with respect to θ of the expectation $y(\theta) = \mathbb{E}(Y(\theta))$. We can approximate this derivative by finite differences. To understand existing possibilities let us take a function $f(\theta)$ of class C^3. We can approximate its derivative in at least two ways:

$$f'(\theta) \approx h^{-1}\big(f(\theta+h) - f(\theta)\big), \quad \text{forward difference},$$
$$f'(\theta) \approx (2h)^{-1}\big(f(\theta+h) - f(\theta-h)\big), \quad \text{central difference}.$$

Each of these methods gives a different error

$$\Big|\frac{f(\theta+h)-f(\theta)}{h} - f'(\theta)\Big| = O(h),$$
$$\Big|\frac{f(\theta+h)-f(\theta-h)}{2h} - f'(\theta)\Big| = O(h^2).$$

Take N simulations $Y_j(\theta)$, $j = 1, \ldots, N$. Let $\hat{y}(\theta) = \frac{1}{N}\sum_{j=1}^{N} Y_j(\theta)$ be the estimator of $y(\theta)$. We can estimate $y'(\theta)$ using forward or central differences

$$\hat{y}'_F(\theta) = \frac{1}{N}\sum_{j=1}^{N} h^{-1}\big(Y_j(\theta+h) - Y_j(\theta)\big),$$
$$\hat{y}'_C(\theta) = \frac{1}{N}\sum_{j=1}^{N} (2h)^{-1}\big(Y_j(\theta+h) - Y_j(\theta-h)\big).$$

We also have a choice in simulating $Y(\theta)$: (i) we can simulate $Y(\theta+h)$ and $Y(\theta)$ ($Y(\theta+h)$ and $Y(\theta-h)$, respectively) independently or (ii) simulate both random variables $Y(\theta+h)$ and $Y(\theta)$ ($Y(\theta+h)$ and $Y(\theta-h)$, respectively) using a common sequence of pseudo-random numbers. Hence, we have in fact four estimators: $\hat{y}'_{F,i}$, $\hat{y}'_{F,ii}$, $\hat{y}'_{C,i}$ and $\hat{y}'_{C,ii}$.

To decide which of these estimators apply in computation, let us analyze the bias for the two methods of derivative approximation

$$\mathrm{Bias}(\hat{y}'_F) = \mathbb{E}(\hat{y}'_F - y'(\theta)) = O(h),$$
$$\mathrm{Bias}(\hat{y}'_C) = \mathbb{E}(\hat{y}'_C - y'(\theta)) = O(h^2).$$

The above formulas suggest that the smaller h the better the accuracy. This conclusion is premature. The effect on bias must be mitigated by the effect on variance. Let us compute the variances for the aforementioned methods of simulation $Y(\theta+h)$ and $Y(\theta)$ ($Y(\theta+h)$ and $Y(\theta-h)$, respectively). When the sequence $Y_j(\theta+h)$ is generated independently from the sequence $Y_j(\theta)$, then

$$\begin{aligned}
&\mathrm{Var}\left(\frac{1}{N}\sum_{j=1}^{N}\frac{Y_j(\theta+h) - Y_j(\theta)}{h}\right)\\
&\quad= \frac{1}{h^2}\frac{1}{N^2}\sum_{j=1}^{N}\Big(\mathrm{Var}\big(Y_j(\theta+h)\big) + \mathrm{Var}\big(Y_j(\theta)\big)\Big)\\
&\quad\simeq \frac{1}{h^2}\left(\frac{\mathrm{Var}(Y(\theta+h))}{N} + \frac{\mathrm{Var}(Y(\theta))}{N}\right) \to \frac{2}{h^2 N}\mathrm{Var}(Y(\theta)).
\end{aligned}$$

Hence, $\mathrm{Var}(\hat{y}'_{F,i}) = O(N^{-1}h^{-2})$.

Similar computations for $Y(\theta+h)$ and $Y(\theta)$ ($Y(\theta+h)$ and $Y(\theta-h)$, respectively) simulated from common pseudo-random numbers, when $Y(\theta)$ fulfills the

assumptions of Lemma 3.14 (see the next section), give

$$\mathrm{Var}\Big(\frac{1}{N}\sum_{j=1}^{N}\frac{Y_j(\theta+h)-Y_j(\theta)}{h}\Big) \simeq \frac{1}{N}\mathrm{Var}\Big(\frac{Y(\theta+h)-Y(\theta)}{h}\Big) \to \frac{1}{N}\mathrm{Var}\big(Y'(\theta)\big).$$

Hence $\mathrm{Var}(\hat{y}'_{F,ii}) = O(N^{-1})$.

If the assumptions of Lemma 3.14 are not fulfilled, we can only expect that $\mathrm{Var}\big(Y(\theta+h)-Y(\theta)\big) \to 0$ for $h \to 0$. There are no rigorous results about the rate of convergence. The numerical experience says that the typical rate is $\mathrm{Var}(\hat{y}'_{F,ii}) = O(N^{-1}h^{-1})$. The results for the estimators $\hat{y}'_{C,i}$ and $\hat{y}'_{C,ii}$ are similar. Hence, the variance increases with $h \to 0$. The best choice of the step size h has to be a compromise between a small h which decreases $\mathrm{Bias}(\hat{y}')$ and a large h which decreases the variance of the estimator.

Example. 3.13 *Let us compute the value of Delta for an option with payoff $g(S_T)$ in the Black-Scholes model. As*

$$S_T^x = x\exp\Big(\big(r-\sigma^2/2\big)T+\sigma W_T\Big),$$

we have

$$\Delta = \frac{1}{2hN}\sum_{j=1}^{N}\Big(g\Big((x+h)\exp\Big(\big(r-\sigma^2/2\big)T+\sigma\sqrt{T}\xi_j\Big)\Big) - g\Big((x-h)\exp\Big(\big(r-\sigma^2/2\big)T+\sigma\sqrt{T}\xi_j\Big)\Big)\Big),$$

where we have applied central differences and a common sequence of normal variables ξ_j to simulate Wiener process.

Pathwise differentiation

The idea is based on the equality

$$\frac{d}{d\theta}y(\theta) = \frac{d}{d\theta}\mathbb{E}\big(Y(\theta)\big) = \mathbb{E}\Big(\frac{dY}{d\theta}\Big). \tag{3.4}$$

The conditions for the validity of (3.4) are given in the following lemma.

LEMMA. 3.14 *Assume that $Y(\theta)$ is differentiable a.s. at θ_0 and satisfies a.s. the Lipschitz condition*

$$\big|Y(\theta_1) - Y(\theta_2)\big| \le M|\theta_1 - \theta_2|,$$

for θ_1, θ_2 in a non-random neighborhood of θ_0 and $\mathbb{E}(M) < +\infty$. Then (3.4) *is satisfied at $\theta = \theta_0$.*

Proof. Let $y(\theta) = \mathbb{E}(Y(\theta))$. Then

$$y'(\theta_0) = \lim_{h\to 0} \frac{y(\theta_0 + h) - y(\theta_0)}{h} = \lim_{h\to 0} \mathbb{E}\Big(\frac{Y(\theta_0 + h) - Y(\theta_0)}{h}\Big). \tag{3.5}$$

Since $\big(Y(\theta_0+h)-Y(\theta_0)\big)/h \le M$ this quotient converges to $Y'(\theta_0)$. By the theorem of dominated convergence the right hand side of (3.5) converges to $\mathbb{E}(Y'(\theta_0))$. ■

Example. 3.15 *As previously, we compute Delta of a call option in the Black-Scholes model*

$$\mathbb{E}\big(g(S_T^x)\big) = \exp(-rT)\mathbb{E}\Big(\big(S_T^x - K\big)^+\Big),$$

where

$$S_T^x = x\exp\Big(\big(r - \sigma^2/2\big)T + \sigma\sqrt{T}\xi\Big), \quad \xi \sim \mathcal{N}(0,1).$$

Then

$$\Delta = \mathbb{E}\left(\frac{dg}{dx}\right) = \mathbb{E}\left(\frac{dg}{dS_T^x}\frac{dS_T^x}{dx}\right).$$

By simple computations we get:

$$\frac{dS_T^x}{dx} = \frac{S_T^x}{x}, \quad \frac{dg}{dS_T^x} = \mathrm{e}^{-rT}\frac{d}{dS_T^x}\big(S_T^x - K\big)^+ = \mathrm{e}^{-rT}\begin{cases} 0, & \text{for } S_T^x < K, \\ 1, & \text{for } S_T^x > K. \end{cases}$$

To compute Delta, we ignore that the derivative in the last equality does not exist for $S_T^x = K$ since this is an event with probability 0.

Eventually, we obtain

$$\Delta = \mathbb{E}\left(\frac{dg}{dx}\right) = \mathrm{e}^{-rT}\mathbb{E}\left(\frac{S_T^x}{x}\mathbb{1}_{S_T^x > K}\right).$$

But Gamma cannot be computed by this approach, even for the Black-Scholes model, as the payoff function is not twice differentiable.

The likelihood ratio method

The most generally applicable method for computation of Greeks is the likelihood ratio method. Suppose we have to compute the derivative of $\mathbb{E}(Y(\theta))$ with respect to θ. The key feature is that the dependence on θ is restricted to the measure $\mathbb{P}_\theta(d\omega)$, i.e.

$$y(\theta) = \mathbb{E}(Y(\theta)) = \int_\Omega Y(\omega)\mathbb{P}_\theta(d\omega).$$

The conditions for the interchange of differentiation and integration are given in the following lemma.

LEMMA. 3.16 *Let $\big(\phi_\theta(x)\big)_{\theta\in\Theta}$ be a family of densities on $\mathbb{R}$ such that $\phi_\theta(x)$ is continuously differentiable with respect to θ a.e. $x \in \mathbb{R}$. Then*

$$\frac{d}{d\theta}\int_{\mathbb{R}} g(x)\phi_\theta(x)dx = \int_{\mathbb{R}} g(x)\frac{d\phi_\theta}{d\theta}(x)dx$$

for θ in an open interval $\Theta_0 \subset \Theta$, if and only if $g \in L^q$ and $\big|\frac{d\phi_\theta}{d\theta}(x)\big| \le M(x)$ x-a.s., for each $\theta \in \Theta_0$, and $M \in L^p$, where $1/p + 1/q = 1$.

Proof. Assume that $(\theta-\epsilon, \theta+\epsilon) \subset \Theta_0$. For $|h| < \epsilon$ we have

$$\begin{aligned}\frac{1}{h}\bigg(\int_{\mathbb{R}} g(x)\phi_{\theta+h}(x)dx - \int_{\mathbb{R}} g(x)\phi_\theta(x)dx\bigg)& \\ = \int_{\mathbb{R}} g(x)\frac{\phi_{\theta+h}(x)-\phi_\theta(x)}{h}dx &= \int_{\mathbb{R}} g(x)\phi'_{\theta^*}(x)dx,\end{aligned} \tag{3.6}$$

where $\theta^* \in (\theta - |h|, \theta + |h|)$.

By the theorem assupmtions $|\phi'_{\theta^*}(x)| \le M(x)$ and

$$\bigg|\int_{\mathbb{R}} g(x)\phi'_{\theta^*}(x)dx\bigg| \le \int_{\mathbb{R}} |g(x)M(x)|dx \le \|g\|_{L^q}\|M\|_{L^p} < +\infty.$$

Hence the left hand side of (3.6) is bounded from above and we can pass to the limit for $h \to 0$ by the dominated convergence theorem. ∎

Example. 3.17 *We illustrate the method by computing Delta and Gamma in the Black-Scholes model, i.e., we compute the x-derivatives of $\mathbb{E}(g(S_T^x))$.*

Assume that we know the transition density from x to S_T^x: $\phi(S_T^x) = \phi(x, S_T^x)$. This function is the density of S_T^x. Then we have

$$\mathbb{E}\big(g(S_T^x)\big) = \int_{\mathbb{R}} g(S_T^x)\phi(x, S_T^x)dS_T^x = \int_{\mathbb{R}} g(S)\phi(x,S)dS.$$

We formally differentiate this integral to obtain Delta

$$\begin{aligned}\Delta =&\frac{\partial}{\partial x}\mathbb{E}\big(g(S_T^x)\big) = \int_{\mathbb{R}} g(S)\frac{\partial}{\partial x}\phi(x,S)dS\\ =&\int_{\mathbb{R}} g(S)\frac{\partial \ln\phi}{\partial x}\phi(x,S)dS = \mathbb{E}\Big(g(S)\frac{\partial \ln\phi}{\partial x}\Big).\end{aligned}$$

The second differentiation gives Gamma (ϕ is a smooth function in the Black-Scholes model)

$$\begin{aligned}\Gamma =&\frac{\partial^2}{\partial x^2}\mathbb{E}\big(g(S_T^x)\big) = \int_{\mathbb{R}} g(S)\Big(\frac{\partial^2 \ln\phi}{\partial x^2}\phi(x,S) + \frac{\partial \ln\phi}{\partial x}\frac{\partial\phi}{\partial x}\Big)dS\\ =&\int_{\mathbb{R}} g(S)\bigg(\frac{\partial^2 \ln\phi}{\partial x^2} + \Big(\frac{\partial \ln\phi}{\partial x}\Big)^2\bigg)\phi(x,S)dS\\ =&\mathbb{E}\bigg(g(S)\bigg(\frac{\partial^2 \ln\phi}{\partial x^2} + \Big(\frac{\partial \ln\phi}{\partial x}\Big)^2\bigg)\bigg).\end{aligned}$$

Using the density of S_T^x in the Black-Scholes model

$$\phi(x,S) = \frac{1}{\sqrt{2\pi\sigma^2 T}S}\exp\bigg(-\frac{\Big(\ln S/x - (r-\sigma^2/2)T\Big)^2}{2\sigma^2 T}\bigg)$$

we obtain

$$\begin{aligned}\frac{\partial \ln\phi}{\partial x} &= \frac{\ln S/x - (r-\sigma^2/2)T}{x\sigma^2 T},\\ \frac{\partial^2 \ln\phi}{\partial x^2} &= -\frac{1+\ln S/x - (r-\sigma^2/2)T}{x^2\sigma^2 T}.\end{aligned}$$

Remark. 3.5 *The likelihood ratio method can be used for computation of other Greeks provided we know $\phi(x,S)$, i.e., for many European options.*

Chapter 4

Integration of stochastic differential equations

The computation of $\mathbb{E}(X)$ from a known distribution of X is a rare event. A more common situation is where X is the value at time T of a stochastic process X_t with the dynamics given by a stochastic differential equation. MC simulations can be used to solve that equation and obtain an approximate distribution of X. In what follows, we restrict our presentation to the Itô stochastic equation

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dW_t,\ t > 0, \quad X_0 = x,$$

where W_t is a standard Wiener process.

Solving that equation means solving the integral equation

$$X_t = X_0 + \int_0^t b(s, X_s)ds + \int_0^t \sigma(s, X_s)dW_s, \tag{4.1}$$

where $\int_0^t \sigma(s, X_s)dW_s$ denotes the Itô integral.

We assume that the coefficients $b(t, x)$ and $\sigma(t, x)$ fulfill the assumptions of Theorem 1.2 and equation (4.1) has a unique strong solution.

The numerical integration of (4.1) means, in fact, a numerical approximation of the corresponding integrals. The main difficulty is in the approximation of the Itô integral.

4.1 Numerical schemes for stochastic differential equations

We begin our analysis with a discrete interpolation of the one-dimensional Wiener process W_t. To interpolate W_t in the interval $[0, T]$ we divide this interval into N

equal subintervals of length $h = T/N$ with grid points

$$t_n = nh = n\frac{T}{N}, \qquad n = 0, \dots, N.$$

The Wiener process W_t can be approximated in points t_n by the expression

$$W_{t_{n+1}} = W_{t_n} + \Delta W_n, \qquad W_{t_0} = W_0 = 0,$$

where ΔW_n are i.i.d. variables with distribution $\mathcal{N}(0, h)$.

LEMMA. 4.1 *Let W_t^h take values W_{t_n} at points t_n and be linearly interpolated between these points. Then*

$$\mathbb{E}\Big(\int_0^T \big|W_t^h - W_t\big|dt\Big) = CN^{-1/2}.$$

Proof. Let $Z_t^T = W_t - \frac{t}{T}W_T$ denote the Brownian bridge in the interval $[0, T]$. Then the process

$$\int_0^T \big|W_t^h - W_t\big|dt$$

has the same distribution as the sum of N copies of

$$\int_0^{T/N} \big|Z_t^{T/N}\big|dt.$$

Let us notice that $\{Z_{tT}^T\}_{0\le t\le 1}$ has the same distribution as $\{\sqrt{T}Z_t^1\}_{0\le t\le 1}$ by the scaling property of Wiener process. Farther, $Z_t^1 \sim \mathcal{N}(0, t(1-t))$ since $\mathbb{E}Z_t^1 = 0$ and $\mathbb{E}\Big(\big(Z_t^1\big)^2\Big) = \mathbb{E}(W_t^2 - 2tW_tW_1 + t^2W_1^2) = t - t^2$.

Hence

$$\begin{aligned}\mathbb{E}\Big(\int_0^T \big|W_t^h - W_t\big|dt\Big) &= N\,\mathbb{E}\Big(\int_0^{T/N} \big|Z_t^{T/N}\big|dt\Big) = T\,\mathbb{E}\Big(\int_0^1 \big|Z_{sT/N}^{T/N}\big|ds\Big)\\ &= \frac{T^{3/2}}{N^{1/2}}\,\mathbb{E}\Big(\int_0^1 \big|Z_s^1\big|ds\Big) = \frac{T^{3/2}}{N^{1/2}}\int_0^1 \sqrt{\frac{2s(1-s)}{\pi}}ds = CN^{-1/2}.\end{aligned}$$

The second equality above is obtained by substitution $t \mapsto s\frac{T}{N}$; the fourth, follows from the formula $\mathbb{E}|\xi| = \sqrt{\frac{2\sigma^2}{\pi}}$ for $\xi \sim \mathcal{N}(0, \sigma^2)$, and the distribution of Z_t^1. ■

Numerical schemes used for the integration of SDEs can be easily obtained by expanding coefficients of the equation by Taylor's formula combined with the Itô lemma and truncating the expansion at an appropriate level. That procedure, called

the *Itô-Taylor expansion*, can be applied to stochastic differential equations of any finite dimension.

Consider the multi-dimensional equation

$$X_t = X_{t_0} + \int_{t_0}^{t} b(s, X_s)ds + \int_{t_0}^{t} \sigma(s, X_s)dW_s, \tag{4.2}$$

where X and b are d-dimensional vectors, σ is a matrix with dimension $d \times m$ and W_t is an m-dimensional Wiener process. Applying the Itô lemma to the coefficients b and σ we obtain

$$\begin{aligned} b(s, X_s) = b(t_0, X_{t_0}) &+ \int_{t_0}^{s} \frac{\partial b(v, X_v)}{\partial v}dv \\ &+ \int_{t_0}^{s} \sum_{i=1}^{d} \frac{\partial b(v, X_v)}{\partial x_i} b_i(v, X_v)dv \\ &+ \frac{1}{2}\int_{t_0}^{s} \sum_{i,j=1}^{d} \frac{\partial^2 b(v, X_v)}{\partial x_i \partial x_j} \left(\sum_{k,l=1}^{m} \sigma_i^k(v, X_v) Q_{kl} \sigma_j^l(v, X_v) \right) dv \\ &+ \int_{t_0}^{s} \sum_{i=1}^{d} \frac{\partial b(v, X_v)}{\partial x_i} \left(\sum_{j=1}^{m} \sigma_i^j(v, X_v) dW_v^j \right), \end{aligned}$$

and

$$\begin{aligned} \sigma(s, X_s) = \sigma(t_0, X_{t_0}) &+ \int_{t_0}^{s} \frac{\partial \sigma(v, X_v)}{\partial v}dv \\ &+ \int_{t_0}^{s} \sum_{i=1}^{d} \frac{\partial \sigma(v, X_v)}{\partial x_i} b_i(v, X_v)dv \\ &+ \frac{1}{2}\int_{t_0}^{s} \sum_{i,j=1}^{d} \frac{\partial^2 \sigma(v, X_v)}{\partial x_i \partial x_j} \left(\sum_{k,l=1}^{m} \sigma_i^k(v, X_v) Q_{kl} \sigma_j^l(v, X_v) \right) dv \\ &+ \int_{t_0}^{s} \sum_{i=1}^{d} \frac{\partial \sigma(v, X_v)}{\partial x_i} \left(\sum_{j=1}^{m} \sigma_i^j(v, X_v) dW_v^j \right), \end{aligned}$$

where Q denotes the covariance matrix of W_t.

Defining the operators

$$\begin{aligned} L^0 f &= \frac{\partial f}{\partial t} + \sum_{i=1}^{d} \frac{\partial f}{\partial x_i} b_i + \frac{1}{2} \sum_{i,j=1}^{d} \frac{\partial^2 f}{\partial x_i \partial x_j} \left(\sum_{k,l=1}^{m} \sigma_i^k Q_{kl} \sigma_j^l \right), \\ L^j f &= \sum_{i=1}^{d} \frac{\partial f}{\partial x_i} \sigma_i^j, \quad j = 1, \ldots, m, \end{aligned}$$

we can write

$$b(s, X_s) = b(t_0, X_{t_0}) + \int_{t_0}^{s} L^0 b(v, X_v) dv + \sum_{j=1}^{m} \int_{t_0}^{s} L^j b(v, X_v) dW_v^j,$$

$$\sigma(s, X_s) = \sigma(t_0, X_{t_0}) + \int_{t_0}^{s} L^0 \sigma(v, X_v) dv + \sum_{j=1}^{m} \int_{t_0}^{s} L^j \sigma(v, X_v) dW_v^j.$$

Inserting these expansions into (4.2) we get

$$X_t = X_{t_0} + b(t_0, X_{t_0})(t - t_0) + \sigma(t_0, X_{t_0})(W_t - W_{t_0}) + r(t), \tag{4.3}$$

where the remainder is

$$\begin{aligned} r(t) = &\int_{t_0}^{t} \int_{t_0}^{s} L^0 b(v, X_v) dv ds + \sum_{j=1}^{m} \int_{t_0}^{t} \int_{t_0}^{s} L^j b(v, X_v) dW_v^j ds \\ &+ \int_{t_0}^{t} \int_{t_0}^{s} L^0 \sigma(v, X_v) dv dW_s + \sum_{j=1}^{m} \int_{t_0}^{t} \int_{t_0}^{s} L^j \sigma(v, X_v) dW_v^j dW_s \end{aligned} \tag{4.4}$$

By discarding the remainder, we obtain the first-order approximation

$$X_t \approx X_{t_0} + b(t_0, X_{t_0})(t - t_0) + \sigma(t_0, X_{t_0})(W_t - W_{t_0})$$

which is the Euler-Maruyama scheme. In one dimension the algorithm of the Euler-Maruyama scheme has the form:

$$\begin{aligned} &X_0 = x, \\ &X_{n+1} = X_n + b(t_n, X_n)h + \sigma(t_n, X_n)\Delta W_n, \end{aligned}$$

where $\Delta W_n = W_{t_{n+1}} - W_{t_n}$.

To obtain a better approximation we can expand terms $L^j \sigma(t, X_t)$ using the Itô lemma

$$\begin{aligned} L^{j_0} \sigma(t, X_t) = L^{j_0} \sigma(t_0, X_{t_0}) &+ \int_{t_0}^{t} L^0 L^{j_0} \sigma(v, X_v) dv \\ &+ \sum_{j=1}^{m} \int_{t_0}^{t} L^j L^{j_0} \sigma(v, X_v) dW_v^j. \end{aligned}$$

Substituting these expansions into $r(t)$ we obtain the following expression for X_t

$$\begin{aligned} X_t = &X_{t_0} + b(t_0, X_{t_0})(t - t_0) + \sigma(t_0, X_{t_0})(W_t - W_{t_0}) \\ &+ \sum_{j=1}^{m} L^j \sigma(t_0, X_{t_0}) \int_{t_0}^{t} \int_{t_0}^{s} dW_v^j dW_s + r_1(t), \end{aligned} \tag{4.5}$$

where the new remainder is

$$\begin{aligned} r_1(t) = &\int_{t_0}^{t}\int_{t_0}^{s} L^0 b(v, X_v)dvds + \sum_{j=1}^{m}\int_{t_0}^{t}\int_{t_0}^{s} L^j b(v, X_v)dW_v^j ds \\ &+ \int_{t_0}^{t}\int_{t_0}^{s} L^0\sigma(v, X_v)dvdW_s \\ &+ \sum_{j=1}^{m}\int_{t_0}^{t}\int_{t_0}^{s_1}\int_{t_0}^{s_2} L^0 L^j\sigma(v, X_v)dvdW_{s_2}^j dW_{s_1} \\ &+ \sum_{j_1,j_2=1}^{m}\int_{t_0}^{t}\int_{t_0}^{s_1}\int_{t_0}^{s_2} L^{j_1} L^{j_2}\sigma(v, X_v)dW_v^{j_2} dW_{s_2}^{j_1} dW_{s_1}. \end{aligned}$$

Discarding $r_1(t)$ we obtain the multi-dimensional Milstein scheme. To implement this scheme effectively, we have to compute the iterated Itô integrals

$$\int_{t_0}^{t}\int_{t_0}^{s} dW_v^j dW_s$$

which is a highly non-trivial operation. Hence, we limit our considerations to a one-dimensional case where the integral can be easily computed

$$\int_{t_0}^{t}\int_{t_0}^{s} dW_v dW_s = \int_{t_0}^{t}\big(W_s - W_{t_0}\big)dW_s = \frac{1}{2}\big((W_t - W_{t_0})^2 - (t - t_0)\big).$$

Then we obtain the one-dimensional Milstein scheme:

$$\begin{aligned} &X_0 = x, \\ &X_{n+1} = X_n + b(t_n, X_n)h + \sigma(t_n, X_n)\Delta W_n \\ &\qquad + \frac{1}{2}\frac{\partial\sigma}{\partial x}(t_n, X_n)\sigma(t_n, X_n)\big((\Delta W_n)^2 - h\big). \end{aligned}$$

Below we will concentrate on one-dimensional schemes. To investigate the convergence of numerical schemes we have to compare the stochastic process X_t to its numerical approximation. The problem is that numerical schemes generate only a sequence of random variables X_n. Hence, we extend that sequence to a process defined for all t either by the linear interpolation

$$\tilde{X}_t^h = X_n + \frac{t - t_n}{t_{n+1} - t_n}\big(X_{n+1} - X_n\big) \text{ for } t \in [t_n, t_{n+1})$$

or the piecewise constant interpolation

$$\tilde{X}_t^h = X_n \text{ for } t \in [t_n, t_{n+1}),$$

where the superscript h indicates the dependence on the time increment h.

Despite the difference in these approximations, one obtains the same limits for $N \to \infty$.

DEFINITION. 4.2 *A numerical scheme is strongly convergent of order γ, if for each $h < h_0$*

$$\mathbb{E}\Big(\big|X_T - \tilde{X}_T^h\big|\Big) \le C_T\, h^\gamma.$$

A numerical scheme is weakly convergent of order γ, if for each $h < h_0$

$$\big|\mathbb{E}f(X_T) - \mathbb{E}f(\tilde{X}_T^h)\big| \le C_T^f\, h^\gamma,$$

for every f of class $C^{2(1+\gamma)}$, where f and its derivatives up to order $2(1+\gamma)$ have polynomial growth. The constants C_T, C_T^f depend on the SDE and the parameters indicated as indices (T or T and f).

Remark. 4.1 *The relevant order of convergence is $1/2, 1, 3/2, 2, 5/2, \ldots$, i.e., an integer multiplicity of $1/2$. Then $2(1+\gamma)$ is an integer and $C^{2(1+\gamma)}$ is a usual space of functions continuous together with their derivatives.*

4.2 Proofs of convergence

We will analyze the convergence rate of the Euler-Maruyama and Milstein schemes in the one-dimensional case. We begin with the strong convergence of the Euler-Maruyama scheme.

THEOREM. 4.3 *Consider the stochastic differential equation* (4.1) *with the coefficients that fulfill the following conditions:*

(A1) $|b(t,x) - b(t,y)| + |\sigma(t,x) - \sigma(t,y)| \le K|x-y|$;

(A2) $|b(t,x)| + |\sigma(t,x)| \le K(1+|x|)$;

(A3) $|b(t,x) - b(s,x)| + |\sigma(t,x) - \sigma(s,x)| \le K(1+|x|)\sqrt{t-s}$;

for $x, y \in \mathbb{R}$ and $t, s \in [0,T]$, $t > s$.

The Euler-Maruyama scheme for equation (4.1) *reads*

$$X_{n+1}^h = X_n^h + b(t_n, X_n^h)h + \sigma(t_n, X_n^h)\Delta W_n,$$

where

$$h = \frac{T}{N},\ t_n = nh,\ n = 0, \ldots, N; \quad \Delta W_n = W_{t_{n+1}} - W_{t_n},\ n = 0, \ldots, N-1,$$

and the superscript h in X_n^h indicates the dependence on the time step h.

The obtained sequence of random variables is extended to a process on $[0,T]$ by linear approximation which for $t \in [t_n, t_{n+1})$ is given by the expression

$$\tilde{X}_t^h = X_n^h + \int_{t_n}^t b(t_n, X_n^h)ds + \int_{t_n}^t \sigma(t_n, X_n^h)dW_s,\ n = 0, \ldots, N-1,\ \tilde{X}_T^h = X_N^h.$$

Then

$$\mathbb{E}\Big(\sup_{t\in[0,T]} \big|X_t - \tilde{X}_t^h\big|\Big) = O(h^{1/2}).$$

The solution $\tilde{X}_t^h$ of the above theorem is defined piecewise for $t \in [t_n, t_{n+1})$. To define $\tilde{X}_t^h$ by one equation for all $t \in [0,T]$, we introduce for $t \in [0,T]$ the function

$$\Psi_t = \max\{t_n, n = 0, \ldots, N\colon\ t_n \le t\}.$$

Then $\tilde{X}_{\Psi_t}^h = X_n^h$ for $t \in [t_n, t_{n+1})$ and $\tilde{X}_t^h$ is given by the equation

$$\tilde{X}_t^h = x + \int_0^t b(\Psi_s, \tilde{X}_{\Psi_s}^h)ds + \int_0^t \sigma(\Psi_s, \tilde{X}_{\Psi_s}^h)dW_s. \tag{4.6}$$

The proof of Theorem 4.3 requires a uniform bound of the solution $\tilde{X}_t^h$ that can be obtained with the help of the Burkholder-Davis-Gundy inequality [29].

LEMMA. 4.4 (Burkholder-Davis-Gundy inequality) *Let $\big(M_t\big)_{t\in[u,v]}$ be a continuous square integrable martingale. For each $m > 0$ there exist constants $0 < k_m < K_m < \infty$ such that*

$$k_m\mathbb{E}\Big(\big(\langle M\rangle_v\big)^m\Big) \le \mathbb{E}\Big(\sup_{t\in[u,v]} |M_t|^{2m}\Big) \le K_m\mathbb{E}\Big(\big(\langle M\rangle_v\big)^m\Big),$$

where $\langle M\rangle_v$ denotes the quadratic variation of M.

Then we can prove the following estimate.

LEMMA. 4.5 *Let $\tilde{X}_t^h$ be the Euler-Maruyama approximation defined in Theorem 4.3. Then under assumptions (A1) and (A2) of this theorem for each $p \ge 1$, we get the estimate*

$$\sup_{h<1} \mathbb{E}\Big(\sup_{t\in[0,T]} |\tilde{X}_t^h|^p\Big)^{1/p} < +\infty.$$

Proof. Since $h = \frac{T}{N}$ is now fixed, we omit the superscript h in the proof. Then the equation for $\tilde{X}_t := \tilde{X}_t^h$ reads

$$\tilde{X}_t = x + \int_0^t b(\Psi_s, \tilde{X}_{\Psi_s})ds + \int_0^t \sigma(\Psi_s, \tilde{X}_{\Psi_s})dW_s,$$

From the above equation we have the estimate

$$\begin{aligned}\sup_{s\le t}|\tilde{X}_s|^p \le 3^{p-1}\bigg(|x|^p + \sup_{s\le t}\bigg|\int_0^s b(\Psi_v, \tilde{X}_{\Psi_v})dv\bigg|^p \\ + \sup_{s\le t}\bigg|\int_0^s \sigma(\Psi_v, \tilde{X}_{\Psi_v})dW_v\bigg|^p\bigg),\end{aligned} \tag{4.7}$$

which follows from Jensen's inequality for the convex function $|x|^p$

$$|a_1 + \cdots + a_n|^p \le n^{p-1}\big(|a_1|^p + \cdots + |a_n|^p\big).$$

We estimate separately each term in (4.7). By Jensen's inequality we get

$$\begin{aligned}\bigg|\int_0^s b(\Psi_v, \tilde{X}_{\Psi_v})dv\bigg|^p &= s^p\bigg|\frac{1}{s}\int_0^s b(\Psi_v, \tilde{X}_{\Psi_v})dv\bigg|^p \le s^{p-1}\int_0^s \big|b(\Psi_v, \tilde{X}_{\Psi_v})\big|^p dv \\ &\le s^{p-1}K^p\int_0^s \big(1 + |\tilde{X}_{\Psi_v}|\big)^p dv \le C\Big(1 + \int_0^s |\tilde{X}_{\Psi_v}|^p dv\Big).\end{aligned}$$

For fixed h and $n = 0, \ldots, N$, we have $|\tilde{X}_{t_n}| < +\infty$. Hence, by assumption (A2) of Theorem 4.3, the stochastic integral

$$\bigg(\int_0^t \sigma(\Psi_v, \tilde{X}_{\Psi_v})dW_v\bigg)_{t\le T}$$

is a square integrable martingale. By the Burkholder-Davis-Gundy inequality, assumption (A2) of Theorem 4.3, and Hölder's inequality, we get

$$\begin{aligned}\mathbb{E}\bigg(\sup_{s\le t}\bigg|\int_0^s \sigma(\Psi_v, \tilde{X}_{\Psi_v})dW_v\bigg|^p\bigg) &\le K\,\mathbb{E}\bigg(\Big(\int_0^t \big|\sigma(\Psi_v, \tilde{X}_{\Psi_v})\big|^2 dv\Big)^{p/2}\bigg) \\ &\le K\,\mathbb{E}\bigg(t^{p/2-1}\int_0^t \Big(\big|\sigma(\Psi_v, \tilde{X}_{\Psi_v})\big|^2\Big)^{p/2} dv\bigg) \le C\bigg(1 + \int_0^t \mathbb{E}\big|\tilde{X}_{\Psi_v}\big|^p dv\bigg).\end{aligned}$$

Collecting the estimates we obtain

$$\mathbb{E}\Big(\sup_{s\le t}\big|\tilde{X}_s\big|^p\Big) \le C\bigg(1 + \int_0^t \mathbb{E}\big|\tilde{X}_{\Psi_v}\big|^p dv\bigg) \le C\bigg(1 + \int_0^t \mathbb{E}\Big(\sup_{r\le v}\big|\tilde{X}_r\big|^p\Big)dv\bigg).$$

By Gronwall's inequality, we obtain the desired estimate

$$\mathbb{E}\Big(\sup_{s\le t}|\tilde{X}_s|^p\Big) < +\infty.$$

■

The following technical lemma will be useful in the proof of Theorem 4.3.

LEMMA. 4.6 *Let $\tilde{X}_t^h$ be the Euler-Maruyama approximation defined in Theorem 4.3. Then under assumptions (A1) and (A2) of this theorem for each $p \ge 1$, we have the estimate*

$$\sup_{h<1}\mathbb{E}\Big(\sup_{t\in[0,T]}|\tilde{X}_t^h - \tilde{X}_{\Psi_t}^h|^p\Big) \le Ch^{p/2}.$$

Proof. As in the previous proof, we omit the index h. It is sufficient to prove the estimate in a single interval $[t_n, t_{n+1})$. Since $\tilde{X}_t = \tilde{X}_{\Psi_t}$ at each point t_n then for $t \in [t_n, t_{n+1})$ we get by Jensen's inequality

$$\big|\tilde{X}_t - \tilde{X}_{\Psi_t}\big|^p \le C\bigg(\Big|\int_{t_n}^t b(t_n, X_n)dv\Big|^p + \Big|\int_{t_n}^t \sigma(t_n, X_n)dW_v\Big|^p\bigg).$$

By assumption (A2) and Lemma 4.5 we have

$$\mathbb{E}\bigg(\Big|\int_{t_n}^t b(t_n, X_n)dv\Big|^p\bigg) \le Ch^p\,\mathbb{E}\left(1 + |X_n|^p\right) \le Ch^p.$$

Since the stochastic integral

$$\int_{t_n}^t \sigma(t_n, X_n)dW_v$$

is a square integrable martingale then, similarly like in the proof of Lemma 4.5, we obtain

$$\begin{aligned}\mathbb{E}\bigg(\Big|\int_{t_n}^t \sigma(t_n, X_n)dW_v\Big|^p\bigg) &\le C\mathbb{E}\bigg(\Big(\int_{t_n}^{t_{n+1}} |\sigma(t_n, X_n)|^2 dv\Big)^{p/2}\bigg)\\ &\le Ch^{p/2}\mathbb{E}\big(1 + |X_n|^p\big) \le Ch^{p/2}.\end{aligned}$$

Collecting the above estimates we get

$$\mathbb{E}\Big(\sup_{t\in[t_n,t_{n+1})}|\tilde{X}_t - \tilde{X}_{\Psi_t}|^p\Big) \le Ch^p + Ch^{p/2} \le Ch^{p/2},$$

from where the result follows. ■

Proof of Theorem 4.3. Using function Ψ_t, we can write

$$\tilde{X}_t = x + \int_0^t b(\Psi_s, \tilde{X}_{\Psi_s})ds + \int_0^t \sigma(\Psi_s, \tilde{X}_{\Psi_s})dW_s,$$

where, as previously, we omit the superscript h in $\tilde{X}_t^h$. Then for $t \in [t_n, t_{n+1})$ we get

$$X_t - \tilde{X}_t = R_1(t) + R_2(t),$$

where

$$R_1(t) = \int_0^t b(s, X_s)ds - \int_0^t b(\Psi_s, \tilde{X}_{\Psi_s})ds,$$
$$R_2(t) = \int_0^t \sigma(s, X_s)dW_s - \int_0^t \sigma(\Psi_s, \tilde{X}_{\Psi_s})dW_s.$$

Let

$$z(t) = \mathbb{E}\big(X_t - \tilde{X}_t\big)^2,\ r_1(t) = \mathbb{E}\big(R_1(t)\big)^2,\ r_2(t) = \mathbb{E}\big(R_2(t)\big)^2.$$

Then

$$z(t) = \mathbb{E}\big(R_1(t) + R_2(t)\big)^2 \le 2r_1(t) + 2r_2(t).$$

We now estimate separately $r_1(t)$ and $r_2(t)$.

$$R_1(t) = \int_0^t \Big(b(s, X_s)ds - b(s, \tilde{X}_{\Psi_s})\Big)ds + \int_0^t \Big(b(s, \tilde{X}_{\Psi_s}) - b(\Psi_s, \tilde{X}_{\Psi_s})\Big)ds.$$

For the first integral we get the estimate

$$\begin{aligned}
\mathbb{E}\bigg(\Big(\int_0^t \Big(b(s, X_s) - b(s, \tilde{X}_{\Psi_s})\Big)ds\Big)^2\bigg) &\le \mathbb{E}\bigg(\Big(\int_0^t K\big|X_s - \tilde{X}_{\Psi_s}\big|ds\Big)^2\bigg) \\
&\le 2K^2T\,\mathbb{E}\bigg(\int_0^t \Big(\big(X_s - \tilde{X}_s\big)^2 + \big(\tilde{X}_s - \tilde{X}_{\Psi_s}\big)^2\Big)ds\bigg) \\
&\le 2K^2T\int_0^t z(s)ds + 2K^2T\,\mathbb{E}\Big(\int_0^t \big(\tilde{X}_s - \tilde{X}_{\Psi_s}\big)^2 ds\Big) \\
&\le C\int_0^t z(s)ds + Ch
\end{aligned}$$

as $z(s) = \mathbb{E}\big(X_s - \tilde{X}_s\big)^2$ and $\big(\tilde{X}_s - \tilde{X}_{\Psi_s}\big)^2$ is estimated from Lemma 4.6.

For the second integral we obtain

$$\begin{aligned}
&\mathbb{E}\bigg(\Big(\int_0^t \Big(b(s,\tilde{X}_{\Psi_s}) - b(\Psi_s,\tilde{X}_{\Psi_s})\Big)ds\Big)^2\bigg) \\
&\qquad \le T\,\mathbb{E}\bigg(\int_0^t \Big(b(s,\tilde{X}_{\Psi_s}) - b(\Psi_s,\tilde{X}_{\Psi_s})\Big)^2 ds\bigg) \\
&\qquad \le 2T\,\mathbb{E}\bigg(\int_0^t K^2\big(1+|\tilde{X}_{\Psi_s}|^2\big)(s-\Psi_s)ds\bigg) \\
&\qquad \le 2TK^2(1+M)\sum_{k<n}\bigg(\int_{t_k}^{t_{k+1}}(s-t_k)ds\bigg) \\
&\qquad \le 2TK^2(1+M)\cdot\frac{1}{2}h^2 n \le T^2K^2(1+M)h,
\end{aligned}$$

where $M = \mathbb{E}\big(\sup_{s\le T}|\tilde{X}_s|^2\big) < +\infty$ by Lemma 4.5.

Collecting these estimates and assuming $h \le 1$ we obtain

$$r_1(t) = \mathbb{E}\big(R_1(s)\big)^2 \le C_1 h + C_2\int_0^t z(s)ds$$

We perform similar computations for $r_2(t)$.

$$R_2(t) = \int_0^t \Big(\sigma(s,X_s) - \sigma(s,\tilde{X}_{\Psi_s})\Big)dW_s + \int_0^t \Big(\sigma(s,\tilde{X}_{\Psi_s}) - \sigma(\Psi_s,\tilde{X}_{\Psi_s})\Big)dW_s.$$

For the first integral we get the estimate

$$\begin{aligned}
&\mathbb{E}\bigg(\Big(\int_0^t \Big(\sigma(s,X_s) - \sigma(s,\tilde{X}_{\Psi_s})\Big)dW_s\Big)^2\bigg) \\
&\qquad = \mathbb{E}\bigg(\int_0^t \Big(\sigma(s,X_s) - \sigma(s,\tilde{X}_{\Psi_s})\Big)^2 ds\bigg) \\
&\qquad \le 2K^2\,\mathbb{E}\bigg(\int_0^t \big(X_s - \tilde{X}_s\big)^2 ds\bigg) + 2K^2\,\mathbb{E}\bigg(\int_0^t \big(\tilde{X}_s - \tilde{X}_{\Psi_s}\big)^2 ds\bigg) \\
&\qquad \le C\int_0^t z(s)ds + Ch.
\end{aligned}$$

The estimate for the second integral is as follows

$$\begin{aligned}\mathbb{E}\bigg(\Big(\int_0^t \Big(\sigma(s,\tilde{X}_{\Psi_s})-\sigma(\Psi_s,\tilde{X}_{\Psi_s})\Big)dW_s\Big)^2\bigg)&\\ =\mathbb{E}\bigg(\int_0^t\Big(\sigma(s,\tilde{X}_{\Psi_s})-\sigma(\Psi_s,\tilde{X}_{\Psi_s})\Big)^2 ds\bigg)&\\ \le 2\mathbb{E}\bigg(\int_0^t K^2\big(1+|\tilde{X}_{\Psi_s}|^2\big)(s-\Psi_s)ds\bigg)&\\ \le 2K^2(1+M)\cdot\frac{1}{2}h^2 n=(nh)K^2(1+M)h\le TK^2(1+M)h.&\end{aligned}$$

Collecting the estimates we write

$$r_2(t)=\mathbb{E}\big(R_2(t)\big)^2\le C_1h+C_2\int_0^t z(s)ds.$$

The estimates of r_1 and r_2 give together

$$z(t)\le C_1h+C_2\int_0^t z(s)ds.$$

By Gronwall's inequality, we get

$$z(t)\le C_1h\exp(C_2t).$$

Hence by the Doob maximal inequality we complete the proof

$$\begin{aligned}\mathbb{E}\Big(\sup_{t\in[0,T]}|X_t-\tilde{X}_t|\Big)&\le\bigg(\mathbb{E}\Big(\sup_{t\in[0,T]}|X_t-\tilde{X}_t|\Big)^2\bigg)^{1/2}\\ &\le 2\bigg(\mathbb{E}\Big(|X_T-\tilde{X}_T|^2\Big)\bigg)^{1/2}\le C\big(z(T)\big)^{1/2}\le C\,h^{1/2}.\end{aligned}$$

■

For the one-dimensional Milstein scheme, we have a better rate of convergence. To make the proof simpler, we impose assumptions stronger than required.

THEOREM. 4.7 *Let the coefficients of equation* (4.1) *be of class* $C^{1,2}([0,T]\times\mathbb{R})$. *In addition, these coefficients fulfill the conditions:*

(B1) $|b(t,x)|\le K(1+|x|)$ *and* $|b(t,x)-b(s,x)|\le K(1+|x|)\sqrt{t-s}$;

(B2) $|\sigma(t,x)|\le K$ *and* $|\sigma(t,x)-\sigma(s,x)|\le K\sqrt{t-s}$;

(B3) $\left|\frac{\partial b}{\partial x}(t,x)\right| + \left|\frac{\partial \sigma}{\partial x}(t,x)\right| + \left|\frac{\partial^2 b}{\partial x^2}(t,x)\right| + \left|\frac{\partial^2 \sigma}{\partial x^2}(t,x)\right| \le K;$

(B4) $\left|\frac{\partial b}{\partial t}(t,x)\right| + \left|\frac{\partial \sigma}{\partial t}(t,x)\right| \le K(1+|x|).$

In the notation of Theorem 4.3 the Milstein scheme for equation (4.1) *reads*

$$X^h_{n+1} = X^h_n + b(t_n, X^h_n)h + \sigma(t_n, X^h_n)\Delta W_n + \frac{1}{2}\frac{\partial \sigma}{\partial x}(t_n, X^h_n)\sigma(t_n, X^h_n)\big((\Delta W_n)^2 - h\big).$$

The above sequence of random variables is extended to a process on $[0,T]$ *by linear approximation that for* $t \in [t_n, t_{n+1})$ *is given by the expression*

$$\tilde{X}^h_t = X^h_n + \int_{t_n}^t b(t_n, X^h_n)ds + \int_{t_n}^t \sigma(t_n, X^h_n)dW_s + \int_{t_n}^t \int_{t_n}^s \frac{\partial \sigma}{\partial x}(t_n, X^h_n)\sigma(t_n, X^h_n)dW_v dW_s,$$

for $n = 0, \dots, N-1$ *with* $\tilde{X}^h_T = X^h_N$.

Under the above assumptions, we have the estimate

$$\mathbb{E}\left(\sup_{t\in[0,T]} \left|X_t - \tilde{X}^h_t\right|\right) = O(h).$$

Proof. To simplify the proof, we will investigate the case of time-independent coefficients. Then the derivatives $\frac{\partial f}{\partial x}$ and $\frac{\partial^2 f}{\partial x^2}$ will be denoted by f' and f''.

We rewrite equation (4.2) in a form more convenient for the proof. For $t \in [t_n, t_{n+1})$ we apply the Itô-Taylor expansion (4.3) with $t_0 = t_n$. Iterating this expansion over all subintervals $[t_k, t_{k+1})$ for $k < n$ we obtain

$$X_t = x + \int_0^t b(X_{\Psi_s})ds + \int_0^t \sigma(X_{\Psi_s})dW_s + \int_0^t \int_{\Psi_s}^s \sigma'(X_{\Psi_s})\sigma(X_{\Psi_s})dW_v dW_s + R(t),$$

where the function $\Psi_t = \max\{t_n\colon\ t_n \le t\}$ is used to simplify the notation. The

remainder $R(t)$ is

$$\begin{aligned}R(t) = &\int_0^t \int_{\Psi_s}^s \big(b'(X_v)b(X_v) + \frac{1}{2}b''(X_v)\sigma^2(X_v)\big)dvds \\ &+ \int_0^t \int_{\Psi_s}^s b'(X_v)\sigma(X_v)dW_v ds \\ &+ \int_0^t \int_{\Psi_s}^s \big(b(X_v)\sigma'(X_v) + \frac{1}{2}\sigma''(X_v)\sigma^2(X_v)\big)dvdW_s \\ &+ \int_0^t \int_{\Psi_s}^s \big(\sigma(X_v)\sigma'(X_v) - \sigma(X_{\Psi_s})\sigma'(X_{\Psi_s})\big)dW_v dW_s.\end{aligned}$$

For the Milstein approximation we have

$$\begin{aligned}\tilde{X}_t = x &+ \int_0^t b(\tilde{X}_{\Psi_s})ds + \int_0^t \sigma(\tilde{X}_{\Psi_s})dW_s \\ &+ \int_0^t \int_{\Psi_s}^s \sigma'(\tilde{X}_{\Psi_s})\sigma(\tilde{X}_{\Psi_s})dW_v dW_s.\end{aligned}$$

Then for $t \in [t_n, t_{n+1})$ we get

$$X_t - \tilde{X}_t = A_1(t) + A_2(t) + A_3(t) + R(t),$$

where

$$\begin{aligned}A_1(t) &= \int_0^t \big(b(X_{\Psi_s}) - b(\tilde{X}_{\Psi_s})\big)ds, \\ A_2(t) &= \int_0^t \big(\sigma(X_{\Psi_s}) - \sigma(\tilde{X}_{\Psi_s})\big)dW_s, \\ A_3(t) &= \int_0^t \int_{\Psi_s}^s \big(\sigma'(X_{\Psi_s})\sigma(X_{\Psi_s}) - \sigma'(\tilde{X}_{\Psi_s})\sigma(\tilde{X}_{\Psi_s})\big)dW_v dW_s.\end{aligned}$$

Let

$$\begin{aligned}z(t) &= \sup_{s\le t} \mathbb{E}(X_s - \tilde{X}_s)^2, \quad a_i(t) = \sup_{s\le t} \mathbb{E}(A_i(s))^2,\ i = 1,2,3, \\ r(t) &= \sup_{s\le t} \mathbb{E}(R(s))^2.\end{aligned}$$

Then

$$z(t) = \sup_{s\le t} \mathbb{E}\big(A_1(s) + A_2(s) + A_3(s) + R(s)\big)^2 \le 4\big(a_1(t) + a_2(t) + a_3(t) + r(t)\big).$$

We now estimate separately every term.

Since $b(x)$ and $\sigma(x)$ have bounded first derivatives, both functions are globally Lipschitz continuous. By this Lipschitz continuity and the Cauchy-Schwarz inequality we get for A_1

$$\begin{aligned}\mathbb{E}\big(A_1^2(t)\big) &= \mathbb{E}\left(\left(\int_0^t \big(b(X_{\Psi_s}) - b(\tilde{X}_{\Psi_s})\big)ds\right)^2\right) \\ &\leq K^2 T \int_0^t \mathbb{E}\Big(\big(X_{\Psi_s} - \tilde{X}_{\Psi_s}\big)^2\Big)ds \leq C\int_0^t z(s)ds.\end{aligned}$$

For A_2 we use the Itô isometry of stochastic integrals

$$\begin{aligned}\mathbb{E}\big(A_2^2(t)\big) &= \mathbb{E}\left(\left(\int_0^t \big(\sigma(X_{\Psi_s}) - \sigma(\tilde{X}_{\Psi_s})\big)dW_s\right)^2\right) \\ &= \int_0^t \mathbb{E}\Big(\big(\sigma(X_{\Psi_s}) - \sigma(\tilde{X}_{\Psi_s})\big)^2\Big)ds \\ &\leq K^2 \int_0^t \mathbb{E}\Big(\big(X_{\Psi_s} - \tilde{X}_{\Psi_s}\big)^2\Big)ds \leq C\int_0^t z(s)ds.\end{aligned}$$

To estimate A_3 we use the boundedness of $\sigma(x)$ and the Lipschitz continuity of $\sigma'(x)$ which follows by the boundedness of $\sigma''(x)$. Applying the Itô isometry two times we obtain

$$\begin{aligned}\mathbb{E}\big(A_3^2(t)\big) &= \mathbb{E}\left(\left(\int_0^t \int_{\Psi_s}^s \big(\sigma'(X_{\Psi_s})\sigma(X_{\Psi_s}) - \sigma'(\tilde{X}_{\Psi_s})\sigma(\tilde{X}_{\Psi_s})\big)dW_v dW_s\right)^2\right) \\ &= \int_0^t \mathbb{E}\left(\left(\int_{\Psi_s}^s \big(\sigma'(X_{\Psi_s})\sigma(X_{\Psi_s}) - \sigma'(\tilde{X}_{\Psi_s})\sigma(\tilde{X}_{\Psi_s})\big)dW_v\right)^2\right)ds \\ &= \int_0^t \int_{\Psi_s}^s \mathbb{E}\Big(\big(\sigma'(X_{\Psi_s})\sigma(X_{\Psi_s}) - \sigma'(\tilde{X}_{\Psi_s})\sigma(\tilde{X}_{\Psi_s})\big)^2\Big)dvds \\ &\leq Ch\int_0^t \mathbb{E}\Big(\big(X_{\Psi_s} - \tilde{X}_{\Psi_s}\big)^2\Big)ds \leq C\int_0^t z(s)ds.\end{aligned}$$

We split the remainder $R(t)$ into its four components R_1 to R_4 and estimate each component separately.

By assumptions (B2), (B3), and Theorem 1.2 we get

$$\begin{aligned}\mathbb{E}\big(R_1^2(t)\big) &= \mathbb{E}\left(\left(\int_0^t\int_{\Psi_s}^s \big(b'(X_v)b(X_v)+\frac{1}{2}b''(X_v)\sigma^2(X_v)\big)dvds\right)^2\right)\\ &\le T\,\mathbb{E}\left(\int_0^t\left(\int_{\Psi_s}^s \big(b'(X_v)b(X_v)+\frac{1}{2}b''(X_v)\sigma^2(X_v)\big)dv\right)^2 ds\right)\\ &\le Ch\int_0^t\int_{\Psi_s}^s \mathbb{E}\big(1+|X_v|^2\big)dvds \le Ch^2\big(1+|x|^2\big).\end{aligned}$$

In the estimate of R_3 we apply additionally the Itô isometry

$$\begin{aligned}\mathbb{E}\big(R_3^2(t)\big) &= \mathbb{E}\left(\left(\int_0^t\int_{\Psi_s}^s \big(b(X_v)\sigma'(X_v)+\frac{1}{2}\sigma''(X_v)\sigma^2(X_v)\big)dvdW_s\right)^2\right)\\ &= \int_0^t \mathbb{E}\left(\left(\int_{\Psi_s}^s \big(b'(X_v)b(X_v)+\frac{1}{2}b''(X_v)\sigma^2(X_v)\big)dv\right)^2\right)ds\\ &\le Ch\int_0^t\int_{\Psi_s}^s \mathbb{E}\big(1+|X_v|^2\big)dvds \le Ch^2\big(1+|x|^2\big).\end{aligned}$$

The estimation of R_2 is more complicated. First, we split the domain of integration $[0,t]$ into the subintervals defined by the grid point of the Milstein approximation

$$\begin{aligned}\mathbb{E}\big(R_2^2(t)\big) &= \mathbb{E}\left(\left(\int_0^t\int_{\Psi_s}^s b'(X_v)\sigma(X_v)dW_vds\right)^2\right)\\ &\le 2\,\mathbb{E}\left(\left(\sum_{k=0}^{n_t-1}\int_{t_k}^{t_{k+1}}\int_{t_k}^s b'(X_v)\sigma(X_v)dW_vds\right)^2\right)\\ &\quad + 2\,\mathbb{E}\left(\left(\int_{t_{n_t}}^t\int_{t_{n_t}}^s b'(X_v)\sigma(X_v)dW_vds\right)^2\right),\end{aligned}$$

where $n_t = \max\{n\colon\ t_n \le t\}$.

By Fubini's theorem we get

$$\begin{aligned}\mathbb{E}\big(R_2^2(t)\big) &\le 2\,\mathbb{E}\left(\left(\sum_{k=0}^{n_t-1}\int_{t_k}^{t_{k+1}}\int_{v}^{t_{k+1}} b'(X_v)\sigma(X_v)dsdW_v\right)^2\right)\\ &\quad + 2\,\mathbb{E}\left(\left(\int_{t_{n_t}}^t\int_{v}^{t} b'(X_v)\sigma(X_v)dsdW_v\right)^2\right).\end{aligned}$$

To estimate the first term, let us observe that since the intervals $[t_k, t_{k+1})$ and $[t_j, t_{j+1})$ are disjoint for $k \neq j$ then

$$\mathbb{E}\bigg(\int_{t_k}^{t_{k+1}}(\dots)dW_v \int_{t_j}^{t_{j+1}}(\dots)dW_v\bigg) = 0.$$

Using that identity in computing the square of the sum in the right hand side and applying the Itô isometry we obtain the estimate

$$\begin{aligned}
\mathbb{E}\big(R_2^2(t)\big) &\leq 2\int_0^{t_{n_t}} \mathbb{E}\bigg(\bigg(\int_v^{t_{k+1}} b'(X_v)\sigma(X_v)ds\bigg)^2\bigg)dv \\
&\quad + 2\int_{t_{n_t}}^{t} \mathbb{E}\bigg(\bigg(\int_v^{t} b'(X_v)\sigma(X_v)ds\bigg)^2\bigg)dv \\
&\leq C\int_0^{t} \mathbb{E}\bigg(\bigg(\int_v^{\Psi_v+h} b'(X_v)\sigma(X_v)ds\bigg)^2\bigg)dv \\
&\leq Ch\int_0^t\int_v^{\Psi_v+h} dsdv \leq Ch^2.
\end{aligned}$$

To estimate R_4 we use the Itô isometry twice, assumptions (B2), (B3) and the estimates of Theorem 1.2 to obtain

$$\begin{aligned}
\mathbb{E}\big(R_4^2(t)\big) &= \mathbb{E}\bigg(\bigg(\int_0^t\int_{\Psi_s}^{s}\big(\sigma(X_v)\sigma'(X_v) - \sigma(X_{\Psi_s})\sigma'(X_{\Psi_s})\big)dW_v dW_s\bigg)^2\bigg) \\
&= \int_0^t \mathbb{E}\bigg(\bigg(\int_{\Psi_s}^{s}\big(\sigma(X_v)\sigma'(X_v) - \sigma(X_{\Psi_s})\sigma'(X_{\Psi_s})\big)dW_v\bigg)^2\bigg)ds \\
&= \int_0^t\int_{\Psi_s}^{s} \mathbb{E}\bigg(\big(\sigma(X_v)\sigma'(X_v) - \sigma(X_{\Psi_s})\sigma'(X_{\Psi_s})\big)^2\bigg)dvds \\
&\leq C\int_0^t\int_{\Psi_s}^{s} \mathbb{E}\big((X_v - X_{\Psi_s})^2\big)dvds \\
&\leq C\big(1+|x|^2\big)\int_0^t\int_{\Psi_s}^{s}|v-\Psi_s|dvds \leq Ch^2\big(1+|x|^2\big).
\end{aligned}$$

Collecting the estimates we obtain

$$z(t) \leq C_1h^2 + C_2\int_0^t z(s)ds.$$

By Gronwall's inequality, we have

$$z(t) \leq C_1h^2\exp(C_2t).$$

In a similar way like in the proof of Theorem 4.3 we apply the Doob maximal inequality to conclude the proof

$$\mathbb{E}\Big(\sup_{t\in[0,T]}|X_t-\tilde{X}_t|\Big)\le\left(\mathbb{E}\Big(\sup_{t\in[0,T]}|X_t-\tilde{X}_t|\Big)^2\right)^{1/2}$$
$$\le 2\left(\mathbb{E}\Big(|X_T-\tilde{X}_T|^2\Big)\right)^{1/2}\le C\big(z(T)\big)^{1/2}\le C\,h.$$

■

Remark. 4.2 *The limitation of the presentation to one-dimensional equations is not accidental. For the Euler-Maruyama scheme, the proof of convergence in many dimensions remains similar to the proof of Theorem 4.3. The case of the Milstein scheme is different. As we have mentioned deriving this scheme, in many dimensions we have to evaluate the iterated stochastic integrals*

$$\int_{t_k}^{t_{k+1}}\int_{t_k}^{s} dW_v^j dW_s^i,\quad i,j=1,\dots,m.$$

The simulations of these integrals, even in a simplified case of so-called commutative noise*, have a very high computational complexity making the gain of a better convergence rate problematic.*

We will now investigate the order of weak convergence for the presented numerical schemes. To simplify the presentation we will consider only the equation with time independent coefficients

$$dX_s=b(X_s)ds+\sigma(X_s)dW_s,\ s>t,\quad X_t=x. \tag{4.8}$$

We write $X_s^{t,x}$ to indicate the dependence of solution on initial conditions. By Theorem 1.13, and Remark 1.4, we know that if $b,\sigma\in C^4(\mathbb{R}^d)$ with bounded derivatives and $g\in C^4(\mathbb{R}^d)$ with polynomial growth together with its derivatives, then

$$u(t,x)=\mathbb{E}\Big(g(X_T^{t,x})\Big)$$

is a solution of the Cauchy problem

$$\begin{aligned}&\frac{\partial u}{\partial t}+\sum_{i=1}^d b_i(x)\frac{\partial u}{\partial x_i}+\sum_{i,j=1}^d a_{ij}(x)\frac{\partial^2 u}{\partial x_i\partial x_j}=0,\quad t\in[0,T),x\in\mathbb{R}^d,\\ &u(T,x)=g(x),\end{aligned} \tag{4.9}$$

where $a_{ij}=\frac{1}{2}\sum_{k=1}^m\sigma_i^k\sigma_j^k$, and u is a C^1 function with respect to t and C^4 function with respect to x which together with its x-derivatives grows polynomially in x uniformly in t.

COROLLARY. 4.8 *When X_t is a strong solution of* (4.8) *and u is a solution of class $C^{1,2}([0,T]\times\mathbb{R}^d)$ of* (4.9)*, then $Y_t = u(t, X_t)$ is a local martingale.*

Proof. By Itô's formula we have

$$\begin{aligned} dY_t =& \frac{\partial u}{\partial t}dt + \sum_{i=1}^{d}\frac{\partial u}{\partial x_i}\Big(b_i dt + \sum_{k=1}^{m}\sigma_i^k dW_t^k\Big) + \sum_{i,j=1}^{d} a_{ij}\frac{\partial^2 u}{\partial x_i \partial x_j}dt \\ =& \Big(\frac{\partial u}{\partial t} + \sum_{i=1}^{d} b_i\frac{\partial u}{\partial x_i} + \sum_{i,j=1}^{d} a_{ij}\frac{\partial^2 u}{\partial x_i\partial x_j}\Big)dt + \sum_{i=1}^{d}\sum_{k=1}^{m}\frac{\partial u}{\partial x_i}\sigma_i^k dW_t^k \\ =& \sum_{i=1}^{d}\sum_{k=1}^{m}\frac{\partial u}{\partial x_i}\sigma_i^k dW_t^k, \end{aligned}$$

which shows that Y_t is the Itô integral. ∎

For the Euler-Maruyama scheme, we have the following result.

THEOREM. 4.9 *Let assumptions (A1), (A2) of Theorem 4.3 be fulfilled and $b, \sigma \in C^4(\mathbb{R}^d)$ with polynomial growth together with their derivatives up to order 4. If X_t is a solution of* (4.8) *with $X_0 = x$ then for each $g \in C^4(\mathbb{R}^d)$, which has polynomial growth together with its derivatives, and $T > 0$, there exists a constant C_g such that*

$$\Big|\mathbb{E}\Big(g(X_T)\Big) - \mathbb{E}\Big(g(\tilde{X}_T^h)\Big)\Big| \le C_g\, h.$$

Proof. We will present a partial proof omitting some tedious computations. We assume $d = 1$ to simplify presentation and omit the superscript h in $\tilde{X}^h$.

By Corollary 4.8 we have

$$\mathbb{E}\Big(g(X_T)\Big) = \mathbb{E}\Big(u(T, X_T)\Big) = u(0, x).$$

Then

$$\begin{aligned} \mathbb{E}\Big(g(\tilde{X}_T) - g(X_T)\Big) &= \mathbb{E}\Big(u(T,\tilde{X}_T)\Big) - u(0,x) \\ &= \sum_{n=1}^{N}\mathbb{E}\Big(u(t_n,\tilde{X}_{t_n}) - u(t_{n-1},\tilde{X}_{t_{n-1}})\Big). \end{aligned}$$

To prove the theorem, it is enough to estimate the local error

$$\Big|\mathbb{E}\Big(u(t_n,\tilde{X}_{t_n}) - u(t_{n-1},\tilde{X}_{t_{n-1}})\Big)\Big| \le Ch^2$$

since then

$$\begin{aligned}\Big|\mathbb{E}\Big(g(\tilde{X}_T) - g(X_T)\Big)\Big| &\leq \sum_{n=1}^{N} \Big|\mathbb{E}\Big(u(t_n, \tilde{X}_{t_n}) - u(t_{n-1}, \tilde{X}_{t_{n-1}})\Big)\Big| \\ &\leq \sum_{n=1}^{N} Ch^2 = CTh.\end{aligned}$$

By Corollary 4.8 we obtain ($X_t^{t_0,x_0}$ denotes the strong solution of (4.8) with initial data $X_{t_0} = x_0$)

$$\mathbb{E}\Big(u(t_n, X_{t_n}^{t_{n-1},\tilde{X}_{t_{n-1}}})|\mathcal{F}_{t_{n-1}}\Big) = u(t_{n-1}, \tilde{X}_{t_{n-1}}).$$

Inserting this equality into the expression for the local error we get

$$\begin{aligned}&\mathbb{E}\Big(u(t_n, \tilde{X}_{t_n}) - u(t_{n-1}, \tilde{X}_{t_{n-1}})\Big) \\ &\qquad = \mathbb{E}\Big(\mathbb{E}\Big(u(t_n, \tilde{X}_{t_n}) - u(t_{n-1}, \tilde{X}_{t_{n-1}})|\mathcal{F}_{t_{n-1}}\Big)\Big) \\ &\qquad = \mathbb{E}\Big(\mathbb{E}\Big(u(t_n, \tilde{X}_{t_n}) - u(t_{n-1}, \tilde{X}_{t_{n-1}}) - u(t_n, X_{t_n}^{t_{n-1},\tilde{X}_{t_{n-1}}}) \\ &\qquad\qquad +u(t_{n-1}, \tilde{X}_{t_{n-1}})|\mathcal{F}_{t_{n-1}}\Big)\Big) \\ &\qquad = \mathbb{E}\Big(\mathbb{E}\Big(u(t_n, \tilde{X}_{t_n}) - u(t_n, X_{t_n}^{t_{n-1},\tilde{X}_{t_{n-1}}})|\mathcal{F}_{t_{n-1}}\Big)\Big).\end{aligned}$$

We rewrite the conditional expectation in the above expression as the difference of two terms

$$\begin{aligned}&\mathbb{E}\Big(u(t_n,\tilde{X}_{t_n}) - u(t_n, X_{t_n}^{t_{n-1},\tilde{X}_{t_{n-1}}})|\mathcal{F}_{t_{n-1}}\Big) \\ &\quad =\mathbb{E}\Big(u(t_n, \tilde{X}_{t_n}) - u(t_n, \tilde{X}_{t_{n-1}})|\mathcal{F}_{t_{n-1}}\Big) \\ &\qquad - \mathbb{E}\Big(u(t_n, X_{t_n}^{t_{n-1},\tilde{X}_{t_{n-1}}}) - u(t_n, \tilde{X}_{t_{n-1}})|\mathcal{F}_{t_{n-1}}\Big)\end{aligned}$$

and estimate each of these terms expanding $u(t, x+\Delta x)-u(t, x)$ in Taylor's series

with respect to Δx. Then we obtain

$$\begin{aligned}
&\Big|\mathbb{E}\Big(\mathbb{E}\Big(u(t_n,\tilde{X}_{t_n}) - u(t_n, X_{t_n}^{t_{n-1},\tilde{X}_{t_{n-1}}})|\mathcal{F}_{t_{n-1}}\Big)\Big)\Big| \\
&\quad\le \mathbb{E}\Big(\sum_{k=1}^{3}\frac{1}{k!}\Big|\frac{\partial^k u(t_n,\tilde{X}_{t_{n-1}})}{\partial x^k}\Big| \\
&\qquad\qquad\times\Big|\mathbb{E}\Big(\big(\tilde{X}_{t_n}-\tilde{X}_{t_{n-1}}\big)^k - \big(X_{t_n}^{t_{n-1},\tilde{X}_{t_{n-1}}}-\tilde{X}_{t_{n-1}}\big)^k\big|\mathcal{F}_{t_{n-1}}\Big)\Big| \\
&\qquad+\mathbb{E}\big(|R(\tilde{X}_{t_n})|\,\big|F_{t_{n-1}}\big)+\mathbb{E}\big(|R(X_{t_n}^{t_{n-1},\tilde{X}_{t_{n-1}}})|\,\big|F_{t_{n-1}}\big)\Big),
\end{aligned}$$

where

$$R(Z)=\frac{1}{4!}\frac{\partial^4 u(t_n,\tilde{X}_{t_{n-1}}+\theta(Z)(Z-\tilde{X}_{t_{n-1}}))}{\partial x^4}(Z-\tilde{X}_{t_{n-1}})^4.$$

Since $u(t,x)$ is a C^4 function with respect to x growing polynomially together with all its derivatives, then we can find an even number $2q$ and a number $C>0$ such that for $k=1,2,3$

$$\Big|\frac{\partial^k u(t_n,\tilde{X}_{t_{n-1}})}{\partial x^k}\Big|\le C\big(1+|\tilde{X}_{t_{n-1}}|^{2q}\big).$$

After some modifications this estimate remains valid in point $\tilde{X}_{t_{n-1}}+\theta(Z)(Z-\tilde{X}_{t_{n-1}})$ for the fourth derivative $\frac{\partial^4 u}{\partial x^4}$ appearing in the remainder $R(Z)$.

The expression

$$\mathbb{E}\Big(\big(\tilde{X}_{t_n}-\tilde{X}_{t_{n-1}}\big)^k-\big(X_{t_n}^{t_{n-1},\tilde{X}_{t_{n-1}}}-\tilde{X}_{t_{n-1}}\big)^k\big|\mathcal{F}_{t_{n-1}}\Big) \tag{4.10}$$

is estimated separately for each $k=1,2,3$.

For $k=1$ we have

$$\begin{aligned}
&\Big|\mathbb{E}\Big(\big(\tilde{X}_{t_n}-\tilde{X}_{t_{n-1}}\big)-\big(X_{t_n}^{t_{n-1},\tilde{X}_{t_{n-1}}}-\tilde{X}_{t_{n-1}}\big)\big|\mathcal{F}_{t_{n-1}}\Big)\Big| \\
&\quad=\Big|\mathbb{E}\Big(X_{t_n}^{t_{n-1},\tilde{X}_{t_{n-1}}}-\tilde{X}_{t_n}\big|\mathcal{F}_{t_{n-1}}\Big)\Big|\le\mathbb{E}\Big(|r(t_n)|\big|\mathcal{F}_{t_{n-1}}\Big) \\
&\quad\le\mathbb{E}\Big(\Big|\int_{t_{n-1}}^{t_n}\int_{t_{n-1}}^{s}L^0 b(X_v)dvds\Big|\Big|\mathcal{F}_{t_{n-1}}\Big),
\end{aligned}$$

where $r(t)$ given by (4.4) is the remainder in the Itô-Taylor expansion for the Euler-Maruyama approximation and operator L^0 is defined in Section 4.1.

Since $L^0 b$ has a polynomial growth there are numbers $2q$ and $C > 0$ such that

$$\mathbb{E}\Big(\Big|\int_{t_{n-1}}^{t_n}\int_{t_{n-1}}^{s} L^0 b(X_v)dvds\Big|\Big|\mathcal{F}_{t_{n-1}}\Big) \le C\big(1+|\tilde{X}_{t_{n-1}}|^{2q}\big)h^2.$$

Similarly, we can prove

$$\mathbb{E}\Big(|r(t_n)|^2\big|\mathcal{F}_{t_{n-1}}\Big) \le C\big(1+|\tilde{X}_{t_{n-1}}|^{2q}\big)h^2.$$

The above inequality is sufficient to estimate (4.10) for $k=3$. Applying the Hölder inequality to $x^3-y^3=(x-y)(x^2+xy+y^2)$, using the estimate for $|r(t_n)|^2$, and the observation that in $\big(\tilde{X}_{t_n}-\tilde{X}_{t_{n-1}}\big)$ and $\big(X_{t_n}^{t_{n-1},\tilde{X}_{t_{n-1}}}-\tilde{X}_{t_{n-1}}\big)$ all terms have the order of smallness at least $1/2$ with respect to h, we obtain for $k=3$

$$\Big|\mathbb{E}\Big(\big(\tilde{X}_{t_n}-\tilde{X}_{t_{n-1}}\big)^3-\big(X_{t_n}^{t_{n-1},\tilde{X}_{t_{n-1}}}-\tilde{X}_{t_{n-1}}\big)^3\big|\mathcal{F}_{t_{n-1}}\Big)\Big| \le C\big(1+|\tilde{X}_{t_{n-1}}|^{2q}\big)h^2.$$

To estimate (4.10) for $k=2$ we write

$$\begin{aligned}&\mathbb{E}\Big(\big(\tilde{X}_{t_n}-\tilde{X}_{t_{n-1}}\big)^2-\big(X_{t_n}^{t_{n-1},\tilde{X}_{t_{n-1}}}-\tilde{X}_{t_{n-1}}\big)^2\big|\mathcal{F}_{t_{n-1}}\Big)\\ &\quad=\mathbb{E}\Big(r(t_n)\big(\tilde{X}_{t_n}-\tilde{X}_{t_{n-1}}+X_{t_n}^{t_{n-1},\tilde{X}_{t_{n-1}}}-\tilde{X}_{t_{n-1}}\big)\big|\mathcal{F}_{t_{n-1}}\Big).\end{aligned}$$

Since

$$\begin{aligned}\big(\tilde{X}_{t_n}-\tilde{X}_{t_{n-1}}+X_{t_n}^{t_{n-1},\tilde{X}_{t_{n-1}}}-\tilde{X}_{t_{n-1}}\big)&=\int_{t_{n-1}}^{t_n}\big(b(\tilde{X}_{t_{n-1}})+b(X_s^{t_{n-1},\tilde{X}_{t_{n-1}}})\big)ds\\ &\quad+\int_{t_{n-1}}^{t_n}\big(\sigma(\tilde{X}_{t_{n-1}})+\sigma(X_s^{t_{n-1},\tilde{X}_{t_{n-1}}})\big)dW_s,\end{aligned}$$

we estimate separately

$$\mathbb{E}\Big(r(t_n)\int_{t_{n-1}}^{t_n}\big(b(\tilde{X}_{t_{n-1}})+b(X_s^{t_{n-1},\tilde{X}_{t_{n-1}}})\big)ds\big|\mathcal{F}_{t_{n-1}}\Big) \tag{4.11}$$

and

$$\mathbb{E}\Big(r(t_n)\int_{t_{n-1}}^{t_n}\big(\sigma(\tilde{X}_{t_{n-1}})+\sigma(X_s^{t_{n-1},\tilde{X}_{t_{n-1}}})\big)dW_s\big|\mathcal{F}_{t_{n-1}}\Big). \tag{4.12}$$

Taking into account that all terms in $r(t_n)$ are at least of the first order in h, we obtain for (4.11)

$$\mathbb{E}\Big(r(t_n)\int_{t_{n-1}}^{t_n}\big(b(\tilde{X}_{t_{n-1}})+b(X_s^{t_{n-1},\tilde{X}_{t_{n-1}}})\big)ds\big|\mathcal{F}_{t_{n-1}}\Big) \le C\big(1+|\tilde{X}_{t_{n-1}}|^{2q}\big)h^2.$$

To estimate (4.12) let us recall the remainder $r(t_n)$

$$r(t_n) = \int_{t_{n-1}}^{t_n}\int_{t_{n-1}}^{s} L^0 b(X_v)dvds + \int_{t_{n-1}}^{t_n}\int_{t_{n-1}}^{s} L^1 b(X_v)dW_v ds + \int_{t_{n-1}}^{t_n}\int_{t_{n-1}}^{s} L^0\sigma(X_v)dvdW_s + \int_{t_{n-1}}^{t_n}\int_{t_{n-1}}^{s} L^1\sigma(X_v)dW_v dW_s.$$

Inserting this expansion into (4.12) and considering the order of smallness with respect to h, we can estimate all terms by h^2 except the last term. To estimate the last term

$$\mathbb{E}\Big(\int_{t_{n-1}}^{t_n}\int_{t_{n-1}}^{s} L^1\sigma(X_v)dW_v dW_s \cdot \int_{t_{n-1}}^{t_n}\big(\sigma(\tilde{X}_{t_{n-1}})+\sigma(X_z^{t_{n-1},\tilde{X}_{t_{n-1}}})\big)dW_z\Big|\mathcal{F}_{t_{n-1}}\Big)$$

we rewrite the second integral as a sum

$$\int_{t_{n-1}}^{t_n}\big(\sigma(X_z^{t_{n-1},\tilde{X}_{t_{n-1}}}) - \sigma(\tilde{X}_{t_{n-1}})\big)dW_z + \int_{t_{n-1}}^{t_n} 2\sigma(\tilde{X}_{t_{n-1}})dW_z.$$

Then, the integral

$$\mathbb{E}\Big(\int_{t_{n-1}}^{t_n}\int_{t_{n-1}}^{s} L^1\sigma(X_v)dW_v dW_s \cdot \int_{t_{n-1}}^{t_n}\big(\sigma(X_z^{t_{n-1},\tilde{X}_{t_{n-1}}})-\sigma(\tilde{X}_{t_{n-1}})\big)dW_z\Big|\mathcal{F}_{t_{n-1}}\Big)$$

is of order h^2 due to the Hölder inequality, the Lipschitz property of σ, and the estimate of Theorem 1.2

$$\mathbb{E}\Big(\Big|X_z^{t_{n-1},\tilde{X}_{t_{n-1}}} - X_{t_{n-1}}^{t_{n-1},\tilde{X}_{t_{n-1}}}\Big|^2\Big) \le C|z-t_{n-1}|.$$

The estimate is completed observing that

$$\begin{aligned}\mathbb{E}\Big(\int_{t_{n-1}}^{t_n}\int_{t_{n-1}}^{s} L^1\sigma(X_v)dW_v dW_s \cdot \int_{t_{n-1}}^{t_n} 2\sigma(\tilde{X}_{t_{n-1}})dW_s\Big|\mathcal{F}_{t_{n-1}}\Big)&\\ = \mathbb{E}\Big(\int_{t_{n-1}}^{t_n} 2\sigma(\tilde{X}_{t_{n-1}})\int_{t_{n-1}}^{s} L^1\sigma(X_v)dW_v ds\Big|\mathcal{F}_{t_{n-1}}\Big)&\\ = \mathbb{E}\Big(\int_{t_{n-1}}^{t_n}\int_{v}^{t_n} 2\sigma(\tilde{X}_{t_{n-1}})L^1\sigma(X_v)dsdW_v\Big|\mathcal{F}_{t_{n-1}}\Big) = 0.&\end{aligned}$$

Hence we have proved that for $k = 1, 2, 3$

$$\Big|\mathbb{E}\Big(\big(\tilde{X}_{t_n} - \tilde{X}_{t_{n-1}}\big)^k - \big(X_{t_n}^{t_{n-1},\tilde{X}_{t_{n-1}}} - \tilde{X}_{t_{n-1}}\big)^k\Big|\mathcal{F}_{t_{n-1}}\Big)\Big| \le C\big(1+|\tilde{X}_{t_{n-1}}|^{2q}\big)h^2.$$

Since the terms $\big(\tilde{X}_{t_n} - \tilde{X}_{t_{n-1}}\big)$ and $\big(X_{t_n}^{t_{n-1},\tilde{X}_{t_{n-1}}} - \tilde{X}_{t_{n-1}}\big)$ are of order $1/2$ with respect to h, as we have mentioned above, then we obtain for the remainders

$$\mathbb{E}\big(|R(\tilde{X}_{t_n})|\,\big|F_{t_{n-1}}\big) = C\big(1+|\tilde{X}_{t_{n-1}}|^{2q}\big)h^2,$$
$$\mathbb{E}\big(|R(X_{t_n}^{t_{n-1},\tilde{X}_{t_{n-1}}})|\,\big|F_{t_{n-1}}\big) = C\big(1+|\tilde{X}_{t_{n-1}}|^{2q}\big)h^2.$$

Collecting all the above estimates, we obtain the desired estimate of the local error

$$\Big|\mathbb{E}\Big(\mathbb{E}\Big(u(t_n,\tilde{X}_{t_n}) - u(t_n,X_{t_n}^{t_{n-1},\tilde{X}_{t_{n-1}}})|\mathcal{F}_{t_{n-1}}\Big)\Big)\Big|$$
$$\leq C\mathbb{E}\Big(\big(1+|\tilde{X}_{t_{n-1}}|^{2q}\big)\big(1+|\tilde{X}_{t_{n-1}}|^{2q}\big)\Big)h^2 \leq C\big(1+|x|^{4q}\big)h^2,$$

where x is the deterministic initial condition $X_{t_0} = x$, and the last estimate follows from Theorem 1.2. ■

For the Milstein scheme, the order of weak convergence is also equal to 1. The proof is analogous to the proof of the above theorem. The same order of convergence as for the Euler-Maruyama scheme follows from the fact that for the remainder in the Milstein approximation we have $\mathbb{E}\big(|r_1(t_n)|\big|\mathcal{F}_{t_{n-1}}\big) = O(h^2)$ analogously as for the remainder in the Euler-Maruyama approximation.

Chapter 5

Introduction to elliptic and parabolic equations

5.1 Sobolev spaces

Partial differential equations require certain smoothness of their solutions to make the equations meaningful. It is difficult to achieve that goal if derivatives are understood in the classical sense. A space suitable for the analysis of equations with non-smooth solutions is a space of functions with derivatives defined in a weak sense called Sobolev's space. (There are many books on the theory of partial differential equations in Sobolev's spaces. Our presentation follows the book by Evans [19] where the reader can find more complete proofs.) To define Sobolev's spaces, we start with weakening the notion of derivatives. In what follows, we will consider functions defined on U, an open subset of $\mathbb{R}^d$.

DEFINITION. 5.1 *Let $u, v \in L^1_{loc}(U)$ and $\alpha = (\alpha_1, \ldots, \alpha_d)$ be a multi-index. We call v the weak derivative of order α of u and write*

$$D^\alpha u = v,$$

if for each $\varphi \in C_0^\infty(U)$ (smooth with compact support) the equality holds

$$\int_U u(x) D^\alpha \varphi(x) dx = (-1)^{|\alpha|} \int_U v(x)\varphi(x) dx,$$

where $D^\alpha \varphi(x) = \frac{\partial^{|\alpha|}\varphi(x)}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}$ and $|\alpha| = \alpha_1 + \cdots + \alpha_d$.
By Du we denote the weak gradient of u, i.e., the vector $(\frac{\partial u}{\partial x_1}, \ldots, \frac{\partial u}{\partial x_d})$ with the partial derivatives understood in the weak sense.

The Sobolev space is a space of functions that are differentiable in the weak sense (their weak derivatives are well defined).

DEFINITION. 5.2 *The Sobolev space $W^{k,p}(U)$ is a space of functions $u\colon U \to \mathbb{R}$ such that u and all weak derivatives $D^\alpha u$, for $|\alpha| \le k$, belong to $L^p(U)$.*

If $p = 2$, we write $H^k(U)$ instead of $W^{k,2}(U)$.

THEOREM. 5.3 *For each integer $k \ge 1$, and $1 \le p \le \infty$ the Sobolev space $W^{k,p}(U)$ is a Banach space with the norm*

$$\|u\|_{W^{k,p}(U)} = \left(\sum_{|\alpha| \le k} \int_U |D^\alpha u|^p dx \right)^{1/p}, \quad 1 \le p < \infty,$$

$$\|u\|_{W^{k,\infty}(U)} = \sum_{|\alpha| \le k} \operatorname*{ess\,sup}_U |D^\alpha u|, \quad p = \infty.$$

The Sobolev space $W^{k,2}(U) \equiv H^k(U)$ is a Hilbert space.

DEFINITION. 5.4 *$W_0^{k,p}(U)$ is the closure of $C_0^\infty(U)$ in $W^{k,p}(U)$.*

THEOREM. 5.5 (Properties of Sobolev's spaces) *Let $u, v \in W^{k,p}(U)$ and $|\alpha| \le k$. Then*

1. *$D^\alpha u \in W^{k-|\alpha|,p}(U)$ and $D^\beta(D^\alpha u) = D^\alpha(D^\beta u) = D^{\alpha+\beta}u$ for all multi-indices α, β such that $|\alpha| + |\beta| \le k$.*
2. *For all $a_1, a_2 \in \mathbb{R}$ the linear combination $a_1 u + a_2 v \in W^{k,p}(U)$ and $D^\alpha(a_1 u + a_2 v) = a_1 D^\alpha u + a_2 D^\alpha v$.*
3. *If V is an open subset of U, then $u\big|_V \in W^{k,p}(V)$.*
4. *If $\varphi \in C_0^\infty(U)$, then $\varphi u \in W^{k,p}(U)$, and*

$$D^\alpha(\varphi u) = \sum_{\beta \le \alpha} \binom{\alpha}{\beta} D^\beta \varphi D^{\alpha-\beta} u.$$

It appears that although functions from Sobolev's spaces are not smooth, they can be approximated by smooth functions. The theorem below describes a procedure of such approximation.

THEOREM. 5.6 *Let U be an open, bounded set in $\mathbb{R}^d$ with ∂U of class C^1. If $u \in W^{k,p}(U)$ for $1 \le p < \infty$, then there exists a sequence $u_m \in C^\infty(\bar{U})$ converging to u in the norm of $W^{k,p}(U)$.*

The space H^{-1}

The space $H^{-1}(U)$ is the dual space to $H_0^1(U)$, i.e., a space of continuous linear functionals on $H_0^1(U)$. The duality product $\langle f, u\rangle$ between $H^{-1}(U)$ and $H_0^1(U)$ denotes the action of functional $f \in H^{-1}(U)$ on element $u \in H_0^1(U)$. The norm in $H^{-1}(U)$ is given by the formula

$$\|f\|_{H^{-1}(U)} = \sup\Big\{\langle f, u\rangle\colon\ u \in H_0^1(U), \|u\|_{H_0^1(U)} \le 1\Big\}.$$

The following result, which follows straightforwardly from the Riesz theorem, gives the representation of $f \in H^{-1}(U)$.

THEOREM. 5.7 *Let $f \in H^{-1}(U)$. Then there exist functions $f_0, f_1, \ldots, f_d$ belonging to $L^2(U)$ such that for each $u \in H_0^1(U)$*

$$\langle f, u\rangle = \int_U \left(f_0 u + \sum_{i=1}^{d} f_i \frac{\partial u}{\partial x_i}\right) dx.$$

Traces of functions

Since a function in $L^p(U)$ has no value at "point x", there is no natural meaning of "restricting u to ∂U" and formulation of boundary value problems. Analyzing such problems, we have to attach "boundary values" to functions in $L^p(U)$. We begin with an extension of u on a larger set that contains U and also ∂U in its interior. It appears that without difficulty, we can construct an extension on the whole $\mathbb{R}^d$.

THEOREM. 5.8 *Let U be a bounded, open set in $\mathbb{R}^d$ with ∂U of class C^1 compactly embedded in a bounded, open set V ($U \subset\subset V$). Then there exists a bounded linear operator*

$$E\colon\ W^{1,p}(U) \to W^{1,p}(\mathbb{R}^d),$$

such that for each $u \in W^{1,p}(U)$, $1 \le p \le \infty$, we have

1. *$Eu = u$ a.s. on U.*
2. *Eu has support in V.*
3. *$\|Eu\|_{W^{1,p}(\mathbb{R}^d)} \le C\|u\|_{W^{1,p}(U)}$, where the constant C depends only on p, U and V.*

Proof. The proof will be carried on for $1 \leq p < \infty$. Fix $x^0 \in \partial U$ and assume that, in a neighborhood of x^0, ∂U is flat, given by equation $x_d = 0$. Let B be an open ball with center x^0 and radius r such that

$$B^+ = B \cap \{x_d \geq 0\} \subset \bar{U},$$
$$B^- = B \cap \{x_d \leq 0\} \subset \mathbb{R}^d \setminus U.$$

Let assume for a moment that $u \in C^\infty(\bar{U})$ and define a higher-order reflection of u from B^+ to B^-

$$\bar{u}(x) = \begin{cases} u(x) & \text{for } x \in B^+, \\ -3u(x_1, \ldots, x_{d-1}, -x_d) + 4u(x_1, \ldots, x_{d-1}, -\frac{x_d}{2}) & \text{for } x \in B^-. \end{cases}$$

We will show that $\bar{u} \in C^1(B)$. Let $u^- = \bar{u}|_{B^-}$, $u^+ = \bar{u}|_{B^+}$. Then $\frac{\partial u^-}{\partial x_d} = \frac{\partial u^+}{\partial x_d}$ on the hyperplane $\{x_d = 0\}$. Indeed by differentiating u^- we get

$$\frac{\partial u^-}{\partial x_d}(x) = 3\frac{\partial u}{\partial x_d}(x_1, \ldots, x_{d-1}, -x_d) - 2\frac{\partial u}{\partial x_d}(x_1, \ldots, x_{d-1}, -\frac{x_d}{2}).$$

This proves that on $\{x_d = 0\}$, we have the desired equality. Since on that hyperplane we have $u^+ = u^-$, then the derivatives with respect to x_i, $i = 1, \ldots, d-1$ are equal. Hence

$$D^\alpha u^-|_{\{x_d=0\}} = D^\alpha u^+|_{\{x_d=0\}}$$

for $|\alpha| \leq 1$, which proves $\bar{u} \in C^1(B)$.

By the above computations, we also have

$$\|\bar{u}\|_{W^{1,p}(B)} \leq C\|u\|_{W^{1,p}(B^+)}$$

for the constant C independent of u.

This estimate can be extended on an arbitrary boundary of class C^1, as each C^1 boundary can be straightened out near x^0 by a diffeomorphic transformation. Then we get

$$\|\bar{u}\|_{W^{1,p}(F)} \leq C\|u\|_{W^{1,p}(U)},$$

where F is the inverse image of B by this diffeomorphic transformation.

Since ∂U is compact, there are finitely many points x_i^0, their neighborhoods F_i, and extensions $\bar{u}_i$ on F_i such that the sum of F_i covers the whole ∂U. Taking a partition of unity ζ_i associated to this covering of ∂U and defining $\bar{u} = \sum \zeta_i u_i$, we obtain

$$\|\bar{u}\|_{W^{1,p}(\mathbb{R}^d)} \leq C\|u\|_{W^{1,p}(U)}, \tag{5.1}$$

where the constant C depends on p, U and d, but is independent of u. Let us observe that the partition of unity ζ_i can be chosen to lie in a selected set $V \supset\supset U$.

By construction $Eu = \bar{u}$ is a bounded, linear operator for $u \in C^\infty(\bar{U})$. Passage from $u \in C^\infty(\bar{U})$ to $u \in W^{1,p}(U)$ can be obtained by the smooth approximation of function in $W^{1,p}(U)$ (Theorem 5.6). Let $u_m \in C^\infty(\bar{U})$ approximate $u \in W^{1,p}(U)$. By the linearity of E and estimate (5.1) we obtain

$$\|Eu_m - Eu_n\|_{W^{1,p}(\mathbb{R}^d)} \leq C\|u_m - u_n\|_{W^{1,p}(U)}.$$

This shows that Eu_m is a Cauchy sequence converging to $\bar{u} = Eu$. ∎

We can now address the problem of boundary values of u on ∂U. The following theorem explains how to define "boundary values".

THEOREM. 5.9 (Trace theorem) *Let U be an open, bounded set in $\mathbb{R}^d$ with ∂U of class C^1. There exists a linear, bounded operator*

$$T\colon\ W^{1,p}(U) \to L^p(\partial U), \quad 1 \leq p < \infty$$

such that

1. $Tu = u\big|_{\partial U}$ *for* $u \in W^{1,p}(U) \cap C(\bar{U})$.
2. $\|Tu\|_{L^p(\partial U)} \leq C\|u\|_{W^{1,p}(U)}$, *where the constant C depends on p and U.*

Proof. Similarly to the proof of the previous theorem, we select $x^0 \in \partial U$ and assume that in a neighborhood of this point the boundary ∂U is flat ($x_d = 0$). Let B be an open ball with center x^0 and radius r. Let $\hat{B}$ be the concentric ball with radius $r/2$. We select a function $\zeta \in C_0^\infty$ such that $\zeta \geq 0$, $\zeta = 1$ on $\hat{B}$ and $\zeta = 0$ in the exterior of B. Denoting $x' = (x_1, \ldots, x_{d-1}) \in \mathbb{R}^{d-1}$ and $\Gamma = \partial U \cap \hat{B}$ we get for $u \in C^1(\bar{U})$

$$\begin{aligned}
\int_\Gamma |u|^p dx' &\leq \int_{\{x_d=0\}} \zeta|u|^p dx' = -\int_{B^+} \frac{\partial}{\partial x_d}\Big(\zeta|u|^p\Big)dx \\
&= -\int_{B^+}\Big(|u|^p\frac{\partial\zeta}{\partial x_d} + p|u|^{p-1}\operatorname{sgn} u\frac{\partial u}{\partial x_d}\zeta\Big)dx \\
&\leq C\int_{B^+}\big(|u|^p + |Du|^p\big)dx,
\end{aligned}$$

where $B^+ = B \cap \{x_d \geq 0\}$ and the last inequality is due to Young's inequality $ab \leq \frac{a^p}{p} + \frac{b^q}{q}$.

Since ∂U is compact, then similarly to the previous proof, we can cover this set by a finite number of subsets and straighten out the boundary on each subset by a homeomorphism of class C^1. Then we get

$$\|u\|_{L^p(\partial U)} \leq C\|u\|_{W^{1,p}(U)}.$$

If for $u \in W^{1,p}(U) \cap C^1(\bar{U})$ we define $Tu = u|_{\partial U}$, then we can write

$$\|Tu\|_{L^p(\partial U)} \leq C\|u\|_{W^{1,p}(U)}, \quad u \in W^{1,p}(U) \cap C^1(\bar{U}).$$

By Theorem 5.6 we can remove the condition $u \in C^1(\bar{U})$. If $u \in W^{1,p}(U)$, there exists a sequence $u_m \in C^\infty(\bar{U})$ converging to u. By the linearity of T we get

$$\|Tu_m - Tu_n\|_{L^p(\partial U)} \leq C\|u_m - u_n\|_{W^{1,p}(U)}.$$

Hence, Tu_m is a Cauchy sequence and defining

$$Tu = \lim_{m\to\infty} Tu_m$$

we extend T to each $u \in W^{1,p}(U)$ with the estimate

$$\|Tu\|_{L^p(\partial U)} \leq C\|u\|_{W^{1,p}(U)}.$$

■

THEOREM. 5.10 *Let U be a bounded set in $\mathbb{R}^d$ and ∂U be of class C^1. If $u \in W^{1,p}(U)$, then $u \in W_0^{1,p}(U)$ if and only if $Tu = 0$ on ∂U.*

Sobolev inequalities

We are now in the position to prove several inequalities between norms of various Sobolev spaces. These inequalities prove the boundedness of embeddings between different Sobolev spaces, which will serve the characterization of the regularity of solutions of differential equations.

We begin with an embedding for functional spaces defined on the whole $\mathbb{R}^d$.

THEOREM. 5.11 (Gagliardo-Nirenberg-Sobolev inequality) *Let $u \in C_0^1(\mathbb{R}^d)$ and $1 \leq p < d$. The following inequality holds for a constant C depending only on p and d*

$$\|u\|_{L^{p^*}(\mathbb{R}^d)} \leq C\|Du\|_{L^p(\mathbb{R}^d)},$$

where p^* is defined as $\frac{1}{p^*} = \frac{1}{p} - \frac{1}{d}$ and is called the Sobolev conjugate of p.

Proof. We begin with the proof for $p = 1$. To simplify the notation we write $x^i := (x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_d)$.

Since the support of u is compact then

$$u(x) = \int_{-\infty}^{x_i} \frac{\partial u(x^i)}{\partial x_i} dy_i.$$

That gives the estimate

$$|u(x)| \leq \int_{\mathbb{R}} |\frac{\partial u(x^i)}{\partial x_i}| dy_i,$$

which can be extended to

$$|u(x)|^{\frac{d}{d-1}} \leq \prod_{i=1}^{d} \Big(\int_{\mathbb{R}} |\frac{\partial u(x^i)}{\partial x_i}| dy_i \Big)^{\frac{1}{d-1}}. \tag{5.2}$$

Integrating the above inequality with respect to x_1 and applying the generalized Hölder inequality

$$\Big\| \prod_{i=2}^{d} f_i \Big\|_{L^1} \leq \prod_{i=2}^{d} \|f_i\|_{L^{d-1}} \tag{5.3}$$

we obtain

$$\begin{aligned}
\int_{\mathbb{R}} |u(x)|^{\frac{d}{d-1}} dx_1 &\leq \int_{\mathbb{R}} dx_1 \prod_{i=1}^{d} \Big(\int_{\mathbb{R}} |\frac{\partial u(x^i)}{\partial x_i}| dy_i \Big)^{\frac{1}{d-1}} \\
&= \Big(\int_{\mathbb{R}} |\frac{\partial u(x)}{\partial x_1}| dx_1 \Big)^{\frac{1}{d-1}} \int_{\mathbb{R}} dx_1 \prod_{i=2}^{d} \Big(\int_{\mathbb{R}} |\frac{\partial u(x^i)}{\partial x_i}| dy_i \Big)^{\frac{1}{d-1}} \\
&\leq \Big(\int_{\mathbb{R}} |\frac{\partial u(x)}{\partial x_1}| dx_1 \Big)^{\frac{1}{d-1}} \prod_{i=2}^{d} \Big(\int_{\mathbb{R}^2} |\frac{\partial u(x^i)}{\partial x_i}| dx_1 dy_i \Big)^{\frac{1}{d-1}} \\
&= \Big(\int_{\mathbb{R}} |\frac{\partial u(x)}{\partial x_1}| dx_1 \Big)^{\frac{1}{d-1}} \Big(\int_{\mathbb{R}^2} |\frac{\partial u(x)}{\partial x_2}| dx_1 dx_2 \Big)^{\frac{1}{d-1}} \\
&\times \prod_{i=3}^{d} \Big(\int_{\mathbb{R}^2} |\frac{\partial u(x^i)}{\partial x_i}| dx_1 dy_i \Big)^{\frac{1}{d-1}}.
\end{aligned}$$

Integrating with respect to x_2 and applying again inequality (5.3) we get

$$\begin{aligned}\int_{\mathbb{R}^2} |u(x)|^{\frac{d}{d-1}} dx_1 dx_2 &\leq \left(\int_{\mathbb{R}^2} |\frac{\partial u(x)}{\partial x_2}| dx_1 dx_2\right)^{\frac{1}{d-1}} \int_{\mathbb{R}} dx_2 \left(\int_{\mathbb{R}} |\frac{\partial u(x)}{\partial x_1}| dx_1\right)^{\frac{1}{d-1}} \\ &\times \prod_{i=3}^{d} \left(\int_{\mathbb{R}^2} |\frac{\partial u(x^i)}{\partial x_i}| dx_1 dy_i\right)^{\frac{1}{d-1}} \\ &\leq \left(\int_{\mathbb{R}^2} |\frac{\partial u(x)}{\partial x_2}| dx_1 dx_2\right)^{\frac{1}{d-1}} \left(\int_{\mathbb{R}^2} |\frac{\partial u(x)}{\partial x_1}| dx_1 dx_2\right)^{\frac{1}{d-1}} \\ &\times \prod_{i=3}^{d} \left(\int_{\mathbb{R}^3} |\frac{\partial u(x^i)}{\partial x_i}| dx_1 dx_2 dy_i\right)^{\frac{1}{d-1}}.\end{aligned}$$

Iterating these integrations we finally obtain

$$\begin{aligned}\int_{\mathbb{R}^k} |u(x)|^{\frac{d}{d-1}} dx_1 dx_2 \dots dx_k &\leq \prod_{i=1}^{k} \left(\int_{\mathbb{R}^k} |\frac{\partial u(x)}{\partial x_i}| dx_1 dx_2 \dots dx_k\right)^{\frac{1}{d-1}} \\ &\times \prod_{i=k+1}^{d} \left(\int_{\mathbb{R}^{k+1}} |\frac{\partial u(x^i)}{\partial x_i}| dx_1 dx_2 \dots dx_k dy_i\right)^{\frac{1}{d-1}}.\end{aligned}$$

For $k = d$ we have

$$\begin{aligned}\int_{\mathbb{R}^d} |u(x)|^{\frac{d}{d-1}} dx &\leq \prod_{i=1}^{d} \left(\int_{\mathbb{R}^d} |\frac{\partial u(x)}{\partial x_i}| dx_1 dx_2 \dots dx_d\right)^{\frac{1}{d-1}} \\ &\leq \prod_{i=1}^{d} \left(\int_{\mathbb{R}^d} |Du(x)| dx\right)^{\frac{1}{d-1}} = \left(\int_{\mathbb{R}^d} |Du(x)| dx\right)^{\frac{d}{d-1}}.\end{aligned}$$

This proves the theorem for $p = 1$.

The proof for $1 < p < d$ is obtained by substituting $v = |u|^\gamma$, with $\gamma > 1$, in the inequality for $p = 1$. Then

$$\begin{aligned}\left(\int_{\mathbb{R}^d} |u(x)|^{\frac{\gamma d}{d-1}} dx\right)^{\frac{d-1}{d}} &\leq \int_{\mathbb{R}^d} D|u(x)|^\gamma dx = \gamma \int_{\mathbb{R}^d} |u(x)|^{\gamma-1} |Du(x)| dx \\ &\leq \gamma \left(\int_{\mathbb{R}^d} |u(x)|^{(\gamma-1)\frac{p}{p-1}} dx\right)^{\frac{p-1}{p}} \left(\int_{\mathbb{R}^d} |Du(x)|^p dx\right)^{\frac{1}{p}}.\end{aligned}$$

Choosing γ such that $\frac{\gamma d}{d-1} = \frac{(\gamma-1)p}{p-1}$ we get $\gamma = \frac{p(d-1)}{d-p}$, which gives $\frac{\gamma d}{d-1} = \frac{dp}{d-p} = p^*$. Such a choice of γ gives the following form of the last inequality

$$\left(\int_{\mathbb{R}^d} |u(x)|^{p^*} dx\right)^{\frac{1}{p^*}} \leq C \left(\int_{\mathbb{R}^d} |Du(x)|^p dx\right)^{\frac{1}{p}}.$$

■

THEOREM. 5.12 (Poincaré's inequality) *Let U be an open, bounded set in $\mathbb{R}^d$. For $u \in W_0^{1,p}(U)$, $1 \leq p < d$, we have the estimate*

$$\|u\|_{L^q(U)} \leq C\|Du\|_{L^p(U)},$$

where $q \in [1, p^*]$ and the constant C depends on p, q, d and U, but is independent of u.

Proof. We approximate $u \in W_0^{1,p}(U)$ by $u_m \in C_0^\infty(U)$. By Theorem 5.8 we extend u_m on $\mathbb{R}^d$ in such a way that they are 0 on $\mathbb{R}^d \setminus U$. By the Gagliardo-Nirenberg-Sobolev inequality we have $\|u_m\|_{L^{p^*}(\mathbb{R}^d)} \leq C\|Du_m\|_{L^p(\mathbb{R}^d)}$. Passing to the limit with m we get $\|u\|_{L^{p^*}(U)} \leq C\|Du\|_{L^p(U)}$. As U is a bounded set, then $\|u\|_{L^q(U)} \leq C\|u\|_{L^{p^*}(U)}$ for $1 \leq q \leq p^*$. ■

The next theorem proves the embedding of Sobolev's spaces into Hölder's spaces defined below.

DEFINITION. 5.13 *Let U be an open set in $\mathbb{R}^d$. The Hölder space $C^{k,\gamma}(U)$ is a space of functions of class $C^k(U)$ whose derivatives of order k are Hölder continuous. A function u: $U \to \mathbb{R}$ belongs to $C^{k,\gamma}(U)$ if $u \in C^k(U)$ and for each multi-index α such that $|\alpha| = k$*

$$\|D^\alpha u\|_{C^{0,\gamma}(U)} = \sup_{\substack{x,y \in U \\ x \neq y}} \left(\frac{|D^\alpha u(x) - D^\alpha u(y)|}{|x-y|^\gamma} \right) < +\infty.$$

THEOREM. 5.14 (Morrey's inequality) *Let $d < p \leq \infty$. Then for $u \in C^1(\mathbb{R}^d)$ we have the estimate*

$$\|u\|_{C^{0,\gamma}(\mathbb{R}^d)} \leq C\|u\|_{W^{1,p}(\mathbb{R}^d)},$$

where the constant C depends only on p and d, and $\gamma = 1 - d/p$.

Proof. We begin with the proof of the inequality

$$⨍_{B(x,r)} |u(y) - u(x)|dy \leq C \int_{B(x,r)} \frac{|Du(y)|}{|y-x|^{d-1}} dy, \tag{5.4}$$

where $B(x,r)$ is a ball with center x and radius r. By $⨍_{B(x,r)} f(x)dx$ we denote the average of f over the ball $B(x,r)$, i.e., $\int_{B(x,r)} f(x)dx$ divided by the volume of $B(x,r)$.

Let $y = x + t\omega$, where $\omega \in \partial B(0,1)$. For $0 < s < r$ we get

$$u(x + s\omega) - u(x) = \int_0^s Du(x + t\omega) \cdot \omega dt.$$

That gives the inequality

$$\begin{aligned}\int_{\partial B(0,1)} |u(x + s\omega) - u(x)| dS(\omega) &\le \int_0^s \int_{\partial B(0,1)} |Du(x + t\omega)| dS(\omega) dt \\ &= \int_0^s t^{d-1} dt \int_{\partial B(0,1)} \frac{|Du(x + t\omega)|}{|x + t\omega - x|^{d-1}} dS(\omega) \\ &= \int_{B(x,s)} \frac{|Du(y)|}{|y - x|^{d-1}} dy \le \int_{B(x,r)} \frac{|Du(y)|}{|y - x|^{d-1}} dy.\end{aligned}$$

Multiplying that inequality by s^{d-1} and integrating over $[0, r]$ we get

$$\int_{B(x,r)} |u(y) - u(x)| dy \le \frac{r^d}{d} \int_{B(x,r)} \frac{|Du(y)|}{|y - x|^{d-1}} dy.$$

That proves (5.4). We now prove

$$\sup_{\mathbb{R}^d} |u| \le C \|u\|_{W^{1,p}(\mathbb{R}^d)}. \tag{5.5}$$

By inequality (5.4) we get for an arbitrary $x \in \mathbb{R}^d$

$$\begin{aligned}|u(x)| = \fint_{B(x,1)} |u(x)| dy &\le \fint_{B(x,1)} \big(|u(x) - u(y)| + |u(y)|\big) dy \\ &\le C \int_{B(x,1)} \frac{|Du(y)|}{|y - x|^{d-1}} dy + \fint_{B(x,1)} |u(y)| dy \\ &\le C \int_{B(x,1)} \frac{|Du(y)|}{|y - x|^{d-1}} dy + C \|u\|_{L^p(B(x,1))} \\ &\le C \bigg(\int_{\mathbb{R}^d} |Du(y)|^p dy \bigg)^{\frac{1}{p}} \bigg(\int_{B(x,1)} \frac{dy}{|y - x|^{\frac{(d-1)p}{p-1}}} \bigg)^{\frac{p-1}{p}} + C \|u\|_{L^p(B(x,1))} \\ &\le C \|u\|_{W^{1,p}(\mathbb{R}^d)}.\end{aligned}$$

The convergence of the integral

$$\int_{B(x,1)} \frac{dy}{|y - x|^{\frac{(d-1)p}{p-1}}}$$

follows from the inequality $\frac{(d-1)p}{p-1} < d$ valid for $p > d$. As x is arbitrary, that proves (5.5).

Let us choose any two points $x, y \in \mathbb{R}^d$. Take $r = |x-y|$ and $V = B(x,r) \cap B(y,r)$. Then

$$|u(x) - u(y)| \le \fint_V |u(x) - u(z)|dz + \fint_V |u(y) - u(z)|dz.$$

For the integrals on the right hand side we have the estimate

$$\begin{aligned}\fint_V |u(x) - u(z)|dz &\le C \fint_{B(x,r)} |u(x) - u(z)|dz \\ &\le C\left(\int_{B(x,r)} |Du(z)|^p dz\right)^{\frac{1}{p}} \left(\int_{B(x,r)} \frac{dz}{|z-x|^{\frac{(d-1)p}{p-1}}}\right)^{\frac{p-1}{p}} \\ &\le C\left(r^{d-\frac{(d-1)p}{p-1}}\right)^{\frac{p-1}{p}} \|Du\|_{L^p(\mathbb{R}^d)} = Cr^{1-\frac{d}{p}}\|Du\|_{L^p(\mathbb{R}^d)}.\end{aligned}$$

These estimates give together

$$|u(x) - u(y)| \le Cr^{1-\frac{d}{p}}\|Du\|_{L^p(\mathbb{R}^d)} = C|x-y|^{1-\frac{d}{p}}\|Du\|_{L^p(\mathbb{R}^d)}.$$

Hence

$$\|u\|_{C^{0,1-d/p}(\mathbb{R}^d)} = \sup_{x \ne y} \frac{|u(x) - u(y)|}{|x-y|^{1-d/p}} \le C\|Du\|_{L^p(\mathbb{R}^d)}.$$

Together with inequality (5.5) that proves the theorem. ∎

THEOREM. 5.15 (Sobolev inequalities) *Let U be an open, bounded set in $\mathbb{R}^d$ with a C^1 boundary, or the whole $\mathbb{R}^d$. For $u \in W^{k,p}(U)$, $1 \le p < \infty$, we have*

1. *If $k < \frac{d}{p}$ and $l \le k$, then $u \in W^{k-l,q}(U)$, where $\frac{1}{q} = \frac{1}{p} - \frac{l}{d}$. In particular, for $l = k$ we have $u \in L^q(U)$. In addition, for a constant C depending only on k, l, p, d and U we have the estimate*

$$\|u\|_{W^{k-l,q}(U)} \le C\|u\|_{W^{k,p}(U)}.$$

2. *If $k > \frac{d}{p}$, then $u \in C^{k-[d/p]-1,\gamma}(\bar{U})$. In addition, for a constant C depending only on k, p, d, γ and U we have the estimate*

$$\|u\|_{C^{k-[d/p]-1,\gamma}(\bar{U})} \le C\|u\|_{W^{k,p}(U)}.$$

The constant γ is given by the formula

$$\gamma = \begin{cases} \left[\frac{d}{p}\right] + 1 - \frac{d}{p}, & \text{if } d/p \text{ is not an integer,} \\ \text{any positive number} < 1, & \text{if } d/p \text{ is an integer,} \end{cases}$$

where $[a]$ denotes the integer part of a.

Proof. Let $k < \frac{d}{p}$. As $D^\alpha u \in L^p(U)$ for $|\alpha| = k$, then

$$\|D^\beta u\|_{L^{p^*}(U)} \leq C\|u\|_{W^{k,p}(U)}, \quad |\beta| = k-1$$

by the Gagliardo-Nirenberg-Sobolev inequality. Hence, $u \in W^{k-1,p^*}(U)$ and similarly $u \in W^{k-2,p^{**}}(U)$, where $\frac{1}{p^{**}} = \frac{1}{p^*} - \frac{1}{d} = \frac{1}{p} - \frac{2}{d}$. Iterating these embeddings we get after l steps $u \in W^{k-l,q}(U)$ for $\frac{1}{q} = \frac{1}{p} - \frac{l}{d}$. That proves point 1. of the theorem.

Let now $k > \frac{d}{p}$. If d/p is not an integer, then by a similar reasoning as above we get $u \in W^{k-l,r}(U)$ for $\frac{1}{r} = \frac{1}{p} - \frac{l}{d}$ and $lp < d$. Choosing $l = \left[\frac{d}{p}\right]$ we obtain $r > \frac{dp}{d-pl} > d$. Then by Morrey's inequality $D^\alpha u \in C^{0,1-d/r}(\bar{U})$ for $|\alpha| \leq k-l-1$. Since $1 - \frac{d}{r} = 1 - \frac{d}{p} + l = \left[\frac{d}{p}\right] + 1 - \frac{d}{p} = \gamma$, where γ is the exponent in the theorem assertion. Hence, $u \in C^{k-[d/p]-1,\gamma}(\bar{U})$.

If d/p is an integer and $k > \frac{d}{p}$, then we can take $l = \frac{d}{p} - 1$. Similarly as before $u \in W^{k-l,r}(U)$, but this time $r = \frac{dp}{d-pl} = d$. By the Gagliardo-Nirenberg-Sobolev inequality we have $D^\alpha u \in L^q(U)$ for all $d \leq q < \infty$ and $|\alpha| \leq k-l-1 = k - \frac{d}{p}$. Then by Morrey's inequality $D^\alpha u \in C^{0,1-d/q}(\bar{U})$ for $d < q < \infty$ and $|\alpha| \leq k - \frac{d}{p} - 1$. It follows then $u \in C^{k-d/p-1,\gamma}(\bar{U})$ for each $0 < \gamma < 1$. ∎

For functions depending on the time variable, embedding theorems can be more complicated as the function can belong to one functional space and its time derivative to another space. We present without proof a theorem that will be used in the analysis of parabolic equations.

THEOREM. 5.16 *Let $u \in L^2(0,T;H_0^1(U))$, $\frac{du}{dt} \in L^2(0,T;H^{-1}(U))$. Then $u \in C([0,T];L^2(U))$ and we have the estimate*

$$\max_{0\leq t\leq T} \|u(t)\|_{L^2(U)} \leq C\Big(\|u\|_{L^2(0,T;H_0^1(U))} + \|\tfrac{du}{dt}\|_{L^2(0,T;H^{-1}(U))}\Big),$$

where the constant C depends only on T.

5.2 Elliptic equations of second-order

We will now study the existence of solutions to elliptic differential equations in an open, bounded set $U \subset \mathbb{R}^d$. The condition that the trace of a solution on the boundary ∂U is equal to a given function is called the Dirichlet boundary condition and the differential problem with the Dirichlet boundary condition is called the Dirichlet problem. We start with the investigation of the Dirichlet problem with zero boundary conditions

$$\begin{aligned} &\mathcal{A}u = f, \quad \text{in } U, \\ &u|_{\partial U} = 0. \end{aligned} \tag{5.6}$$

$\mathcal{A}$ is a second-order differential operator. Depending on the situation that operator is considered in the *divergence* form

$$\mathcal{A}u(x) = \sum_{i=1}^{d}\sum_{j=1}^{d} -\frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial u(x)}{\partial x_j}\right) + \sum_{i=1}^{d} b_i(x)\frac{\partial u(x)}{\partial x_i} + c(x)u(x) \tag{5.7}$$

or the *nondivergence* form

$$\mathcal{A}u(x) = \sum_{i=1}^{d}\sum_{j=1}^{d} -a_{ij}(x)\frac{\partial^2 u(x)}{\partial x_i \partial x_j} + \sum_{i=1}^{d} b_i(x)\frac{\partial u(x)}{\partial x_i} + c(x)u(x). \tag{5.8}$$

Remark. 5.1 *If the coefficients a_{ij} are of class C^1, then an operator written in divergence form can be rewritten in nondivergence form with b_i replaced by*

$$\bar{b}_i = b_i - \sum_{j=1}^{d} \frac{\partial a_{ij}}{\partial x_j}.$$

We will assume that $\mathcal{A}$ is a uniformly elliptic operator.

DEFINITION. 5.17 *The differential operator $\mathcal{A}$ defined by* (5.7) *or* (5.8) *is called uniformly elliptic, if $a_{ij} = a_{ji}$ $(i, j = 1, \ldots, d)$ and*

$$\exists \delta > 0 \; \forall x \in U \; \forall \xi \in \mathbb{R}^d \setminus \{0\} \quad \sum_{i=1}^{d}\sum_{j=1}^{d} a_{ij}(x)\xi_i\xi_j \geq \delta \|\xi\|^2.$$

Our goal is to prove the existence (and uniqueness) of solutions for the Dirichlet problem with a uniformly elliptic operator $\mathcal{A}$. For the Dirichlet problem (5.6), we define a weak solution.

DEFINITION. 5.18 *Let a_{ij}, b_i, $c \in L^\infty(U)$ $(i,j = 1,\dots,d)$ and $f \in L^2(U)$. We call $u \in H_0^1(U)$ a weak solution of the Dirichlet problem* (5.6) *with $\mathcal{A}$ given in divergence form if*

$$\int_U \Big(\sum_{i,j=1}^d a_{ij}(x) \frac{\partial u(x)}{\partial x_i} \frac{\partial v(x)}{\partial x_j} + \sum_{i=1}^d b_i(x) \frac{\partial u(x)}{\partial x_i} v(x) + c(x)u(x)v(x) \Big) dx = \int_U f(x)v(x)dx,$$

for each $v \in H_0^1(U)$

The existence of weak solutions to problem (5.6) follows from the Lax-Milgram theorem.

THEOREM. 5.19 (Lax-Milgram theorem) *Let $B\colon\ H \times H \to \mathbb{R}$ be a bilinear functional in a Hilbert space H such that*

$$\forall u, v \in H \quad |B[u,v]| \le \alpha \|u\| \|v\|,\ \alpha > 0,$$
$$\forall u \in H \quad |B[u,u]| \ge \beta \|u\|^2,\ \beta > 0.$$

If $f\colon\ H \to \mathbb{R}$ is a bounded linear functional in H, then there exists a unique element $u \in H$ such that

$$B[u,v] = (f,v), \quad \forall v \in H,$$

where $\|\cdot\|$ and $(\cdot,\cdot)$ denote the norm and the scalar product in H, respectively.

THEOREM. 5.20 (Energy estimates) *Let $u, v \in H_0^1(U)$ and*

$$B[u,v] = \int_U \Big(\sum_{i,j=1}^d a_{ij} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} + \sum_{i=1}^d b_i \frac{\partial u}{\partial x_i} v + cuv \Big) dx.$$

There exist constants $\alpha, \beta > 0$ and $\gamma \ge 0$ such that

$$|B[u,v]| \le \alpha \|u\|_{H_0^1(U)} \|v\|_{H_0^1(U)},$$
$$\beta \|u\|^2_{H_0^1(U)} \le B[u,u] + \gamma \|u\|^2_{L^2(U)}.$$

Proof. Since a_{ij}, b_i, $c \in L^\infty(U)$ $(i,j = 1,\dots,d)$ we have the estimate

$$|B[u,v]| \le \sum_{i,j=1}^d \|a_{ij}\|_{L^\infty(U)} \int_U |Du||Dv|dx + \sum_{i=1}^d \|b_i\|_{L^\infty(U)} \int_U |Du||v|dx + \|c\|_{L^\infty(U)} \int_U |u||v|dx \le C \|u\|_{H_0^1(U)} \|v\|_{H_0^1(U)}.$$

By the uniform ellipticity

$$\delta \int_U |Du|^2 dx \le \int_U \sum_{i,j=1}^d a_{ij} \frac{\partial u}{\partial x_i} \frac{\partial u}{\partial x_j} dx = B[u,u] - \int_U \Big(\sum_{i=1}^d b_i \frac{\partial u}{\partial x_i} u + cu^2 \Big) dx$$
$$\le B[u,u] + \sum_{i=1}^d \|b_i\|_{L^\infty(U)} \int_U |Du||u| dx + \|c\|_{L^\infty(U)} \int_U |u|^2 dx.$$

In Cauchy's inequality

$$\int_U |Du||u| dx \le \epsilon \int_U |Du|^2 dx + \frac{1}{4\epsilon} \int_U |u|^2 dx$$

we select ϵ such that $\epsilon \sum_{i=1}^d \|b_i\|_{L^\infty(U)} < \frac{\delta}{2}$. Then

$$\frac{\delta}{2} \int_U |Du|^2 dx \le B[u,u] + C \int_U |u|^2 dx.$$

By Poincaré's inequality $\|u\|_{L^2(U)} \le C \|Du\|_{L^2(U)}$ we get the estimate

$$\|u\|_{H_0^1(U)} \le C \|Du\|_{L^2(U)}.$$

It easily follows that

$$\beta \|u\|^2_{H_0^1(U)} \le B[u,u] + \gamma \|u\|^2_{L^2(U)}.$$

■

Remark. 5.2 *In the rest of this chapter (u,v) will denote the standard scalar product in $L^2(U)$ also for $u, v \in H_0^1(U)$.*

THEOREM. 5.21 (Existence of weak solutions) *There is a constant $\gamma \ge 0$ such that for each $\mu \ge \gamma$ and $f \in L^2(U)$, there exists a unique weak solution $u \in H_0^1(U)$ of the Dirichlet problem with $\mathcal{A}$ in divergence form*

$$\begin{aligned} &\mathcal{A}u + \mu u = f, \quad \text{in } U, \\ &u|_{\partial U} = 0. \end{aligned} \tag{5.9}$$

Proof. Let γ be the constant from Theorem 5.20. Taking $\mu \ge \gamma$ we define

$$B_\mu[u,v] = B[u,v] + \mu(u,v), \quad u, v \in H_0^1(U).$$

$B_\mu[u,v]$ fulfills the assumptions of the Lax-Milgram theorem. For $f \in L^2(U)$ we define a linear functional $\langle f, v\rangle = (f, v)$. Applying the Lax-Milgram theorem to the equation

$$B_\mu[u,v] = \langle f, v\rangle,$$

we find a unique $u \in H_0^1(U)$ fulfilling the conclusion of the theorem. Hence u is a unique weak solution of the Dirichlet problem (5.9). ■

We close our analysis of Dirichlet's problem with zero boundary conditions with a theorem that describes the improvement of solution regularity for more regular data. We present this theorem without proof which is technically complicated.

THEOREM. 5.22 (Higher regularity) *Let a_{ij}, b_i, $c \in C^{m+1}(\bar{U})$ $(i,j = 1,\dots,d)$ and $f \in H^m(U)$. Assume that $u \in H_0^1(U)$ is a unique weak solution of the Dirichlet problem* (5.9) *with ∂U of class C^{m+2}. Then $u \in H^{m+2}(U)$ and we have the estimate*

$$\|u\|_{H^{m+2}(U)} \le C\|f\|_{H^m(U)},$$

where the constant C depends only on the coefficients of $\mathcal{A}$, m, and U.

For the Dirichlet problem with non-zero boundary conditions, we have the following theorem.

THEOREM. 5.23 *If the boundary ∂U is of class C^1, then there is a constant $\gamma \ge 0$ such that for each $\mu \ge \gamma$, $f \in L^2(U)$ and $w \in H^1(U)$ there exists a unique weak solution $u \in H^1(U)$ of the Dirichlet problem with $\mathcal{A}$ in divergence form*

$$\begin{aligned} &\mathcal{A}u + \mu u = f, \quad \text{in } U,\\ &Tu = Tw, \quad \text{on } \partial U. \end{aligned} \tag{5.10}$$

Proof. Let $\tilde{u} \in H_0^1(U)$ be a weak solution of the problem

$$\begin{aligned} &\mathcal{A}\tilde{u} + \mu\tilde{u} = \tilde{f}, \quad \text{in } U,\\ &\tilde{u}|_{\partial U} = 0, \end{aligned}$$

where $\tilde{f} = f - \mathcal{A}w - \mu w \in H^{-1}(U)$.

The existence of $\tilde{u}$ follows by the Lax-Milgram theorem, as $\langle \tilde{f}, v\rangle$ defines for $v \in H_0^1(U)$ a bounded linear functional $\langle \tilde{f}, v\rangle = (f,v) - B[w,v] - \mu(w,v)$. Then $u = \tilde{u} + w$ is a unique weak solution of (5.10). ■

5.3 Parabolic equations of second-order

Let $U \subset \mathbb{R}^d$ be an open, bounded domain and $U_T = (0,T] \times U$, $T > 0$. We will study the initial-boundary value problem with the Dirichlet boundary conditions

$$\begin{aligned} &\tfrac{\partial}{\partial t}u + \mathcal{A}^t u = f, \quad \text{in } U_T, \\ &u = 0, \quad \text{on } [0,T] \times \partial U, \\ &u = g, \quad \text{on } \{t=0\} \times U. \end{aligned} \tag{5.11}$$

Similarly as in the elliptic case we consider the operator $\mathcal{A}^t$ in the divergence form

$$\begin{aligned} &\mathcal{A}^t u(t,x) \\ &\quad = \sum_{i=1}^{d}\sum_{j=1}^{d} -\frac{\partial}{\partial x_i}\left(a_{ij}(t,x)\frac{\partial u(t,x)}{\partial x_j}\right) + \sum_{i=1}^{d} b_i(t,x)\frac{\partial u(t,x)}{\partial x_i} + c(t,x)u(t,x) \end{aligned}$$

and the nondivergence form

$$\mathcal{A}^t u(t,x) = \sum_{i=1}^{d}\sum_{j=1}^{d} -a_{ij}(t,x)\frac{\partial^2 u(t,x)}{\partial x_i \partial x_j} + \sum_{i=1}^{d} b_i(t,x)\frac{\partial u(t,x)}{\partial x_i} + c(t,x)u(t,x).$$

DEFINITION. 5.24 *The second-order operator $\frac{\partial}{\partial t} + \mathcal{A}^t$ is called uniformly parabolic if $a_{ij} = a_{ji}$ $(i,j = 1,\dots,d)$ and*

$$\exists \delta > 0 \; \forall (t,x) \in U_T \; \forall \xi \in \mathbb{R}^d \setminus \{0\} \quad \sum_{i=1}^{d}\sum_{j=1}^{d} a_{ij}(t,x)\xi_i\xi_j \geq \delta \|\xi\|^2.$$

The function $u(t,x)$ which solves problem (5.11) will be treated as a mapping $u\colon [0,T] \to H_0^1(U)$. Then $u(t)$ denotes the value of that function in $H_0^1(U)$.

DEFINITION. 5.25 *Let a_{ij}, b_i, $c \in L^\infty(U_T)$ $(i,j = 1,\dots,d)$, $f \in L^2(U_T)$ and $g \in L^2(U)$. A function $u \in L^2(0,T;H_0^1(U))$ is called a weak solution of the initial-boundary value problem* (5.11) *with $\mathcal{A}^t$ in divergence form if $\frac{du}{dt} \in L^2(0,T;H^{-1}(U))$ and*

1. *For any $v \in H_0^1(U)$*

$$\langle \tfrac{d}{dt}u(t), v\rangle + B^t[u(t),v] = (f(t),v),$$

where the equality holds almost everywhere in t and

$$B^t[u,v] = \int_U \left(\sum_{i,j=1}^{d} a_{ij}(t,x)\frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_j} + \sum_{i=1}^{d} b_i(t,x)\frac{\partial u}{\partial x_i}v + c(t,x)uv\right)dx.$$

2. $u(0) = g$.

If $u \in L^2(0,T;H_0^1(U))$ *and* $\frac{du}{dt} \in L^2(0,T;H^{-1}(U))$, *then* $u \in C([0,T];L^2(U))$, *which means that the equality of point 2. makes sens (cf. Theorem 5.16).*

Galerkin approximation

Since $H_0^1(U) \subset L^2(U)$ and both spaces are Hilbert spaces, we can select the basis $\{\psi_k\}_{k=1}^{\infty}$, which is an orthonormal basis of $L^2(U)$ and an orthogonal basis of $H_0^1(U)$. We can choose as ψ_k the basis of eigenvectors in $H_0^1(U)$ of a uniformly elliptic operator $\mathcal{A}_0 = \sum_{i,j=1}^{d} -\frac{\partial}{\partial x_i}\Big(a_{ij}(x)\frac{\partial u(x)}{\partial x_j}\Big)$ (for details see Evans [19][Section 6.5]). For a fixed m, we define the function

$$u_m(t) = \sum_{k=1}^{m} d_m^k(t)\psi_k.$$

We are looking for a choice of coefficients d_m^k which makes u_m an approximation of a solution to (5.11). Thus we seek u_m as a solution of the discrete approximation of (5.11):

$$\begin{aligned} &(\tfrac{d}{dt}u_m(t),\psi_k) + B^t[u_m(t),\psi_k] = (f(t),\psi_k), \quad 0 < t \le T, k = 1,\dots,m,\\ &d_m^k(0) = (g,\psi_k), \quad k = 1,\dots,m. \end{aligned} \tag{5.12}$$

THEOREM. 5.26 *For every* $m \ge 1$ *there exists a unique function* u_m *which solves* (5.12).

Proof. By the definition of u_m we have $(\frac{d}{dt}u_m,\psi_k) = \frac{d}{dt}d_m^k$. We have also the equality

$$B^t[u_m(t),\psi_k] = \sum_{i=1}^{m} e^{ki}(t)d_m^i(t),$$

where $e^{ki}(t) = B^t[\psi_i,\psi_k]$.

Denoting $f_k(t) = (f,\psi_k)$ and using the above equalities we transform (5.12) into a linear system of ordinary differential equations

$$\tfrac{d}{dt}d_m^k + \sum_{i=1}^{m} e^{ki}d_m^i = f_k, \quad k = 1,\dots,m,$$

with initial conditions $d_m^k(0) = (g,\psi_k)$.
As a linear system, it possesses a unique solution. Thus, we obtain a unique solution of (5.12). ∎

THEOREM. 5.27 (Energy estimates) *If u_m is a solution of* (5.12)*, then*

$$\begin{aligned}\max_{0\le t\le T}\|u_m(t)\|_{L^2(U)} + \|u_m\|_{L^2(0,T;H_0^1(U))} + \|\tfrac{d}{dt}u_m\|_{L^2(0,T;H^{-1}(U))}\\ \le C\Big(\|f\|_{L^2(0,T;L^2(U))} + \|g\|_{L^2(U)}\Big), \quad m=1,2,\dots,\end{aligned} \tag{5.13}$$

where the constant C depends on U, T and the coefficients of $\mathcal{A}^t$.

Proof. We estimate separately every term on the left hand side of (5.13).

Multiplying (5.12) by d_m^k and summing over k from 1 to m we get

$$(\tfrac{d}{dt}u_m(t), u_m(t)) + B^t[u_m(t), u_m(t)] = (f(t), u_m(t)). \tag{5.14}$$

By Theorem 5.20 and the uniform boundedness in t of the coefficients of $B^t[u,v]$ we obtain the estimate

$$\beta\|u_m(t)\|^2_{H_0^1(U)} \le B^t[u_m(t), u_m(t)] + \gamma\|u_m(t)\|^2_{L^2(U)},$$

where $\beta > 0$, $\gamma \ge 0$.

Since $(\frac{d}{dt}u_m(t), u_m(t)) = \frac{d}{dt}\big(\frac{1}{2}\|u_m(t)\|^2_{L^2(U)}\big)$ and

$$|(f(t), u_m(t))| \le \frac{1}{2}\|f(t)\|^2_{L^2(U)} + \frac{1}{2}\|u_m(t)\|^2_{L^2(U)},$$

we can use the estimate of Theorem 5.20 to replace (5.14) by the inequality

$$\frac{d}{dt}\|u_m(t)\|^2_{L^2(U)} + 2\beta\|u_m(t)\|^2_{H_0^1(U)} \le C_1\|u_m(t)\|^2_{L^2(U)} + C_2\|f(t)\|^2_{L^2(U)}, \tag{5.15}$$

which holds a.e. in t.

Inequality (5.15) remains valid if we neglect $2\beta\|u_m(t)\|^2_{H_0^1(U)}$ on the left hand side. Then we obtain $\eta' \le C_1\eta + C_2\xi$, where $\eta = \|u_m(t)\|^2_{L^2(U)}$ and $\xi = \|f(t)\|^2_{L^2(U)}$. By Gronwall's lemma we get

$$\eta(t) \le e^{C_1 t}\Big(\eta(0) + C_2\int_0^t \xi(s)ds\Big).$$

Since $\eta(0) = \|u_m(0)\|^2_{L^2(U)} \le \|g\|^2_{L^2(U)}$, then

$$\max_{0\le t\le T}\|u_m(t)\|^2_{L^2(U)} \le C\Big(\|g\|^2_{L^2(U)} + \|f\|^2_{L^2(0,T;L^2(U))}\Big). \tag{5.16}$$

To estimate the second term in (5.13) we integrate (5.15) over $[0,T]$ and apply (5.16)

$$\|u_m\|^2_{L^2(0,T;H^1_0(U))} = \int_0^T \|u_m(t)\|^2_{H^1_0(U)} dt \le C\Big(\|f\|^2_{L^2(0,T;L^2(U))} + \|g\|^2_{L^2(U)}\Big).$$

To estimate the last term of (5.13) we select $v \in H^1_0(U)$, $\|v\|_{H^1_0(U)} \le 1$, and split this v into orthogonal components $v = v_1 + v_2$, where $v_1 \in \mathrm{Lin}(\psi_1, \dots, \psi_m)$ and $(v_2, \psi_k) = 0$, $k = 1, \dots, m$. Then $\|v_1\|_{H^1_0(U)} \le \|v\|_{H^1_0(U)} \le 1$.

From (5.12) we obtain

$$(\tfrac{d}{dt}u_m(t), v_1) + B^t[u_m(t), v_1] = (f(t), v_1).$$

Since

$$\langle \tfrac{d}{dt}u_m(t), v\rangle = (\tfrac{d}{dt}u_m(t), v) = (\tfrac{d}{dt}u_m(t), v_1) = (f(t), v_1) - B^t[u_m(t), v_1],$$

then by the elementary inequality $|a-b| \le |a|+|b|$ and the estimate $\|v_1\|_{H^1_0(U)} \le 1$ we get

$$|\langle \tfrac{d}{dt}u_m(t), v\rangle| \le C\Big(\|f(t)\|_{L^2(U)} + \|u_m(t)\|_{H^1_0(U)}\Big).$$

Thus

$$\|\tfrac{d}{dt}u_m(t)\|_{H^{-1}(U)} \le C\Big(\|f(t)\|_{L^2(U)} + \|u_m(t)\|_{H^1_0(U)}\Big).$$

Squaring the above inequality and integrating we obtain the desired estimate

$$\begin{aligned}\int_0^T \|\tfrac{d}{dt}u_m(t)\|^2_{H^{-1}(U)} dt &\le C\int_0^T \Big(\|f(t)\|^2_{L^2(U)} + \|u_m(t)\|^2_{H^1_0(U)}\Big) dt \\ &\le C\Big(\|f\|^2_{L^2(0,T;L^2(U))} + \|g\|^2_{L^2(U)}\Big).\end{aligned}$$

■

THEOREM. 5.28 *Let a_{ij}, b_i, $c \in L^\infty(U_T)$ $(i,j = 1,\dots,d)$, $f \in L^2(U_T)$ and $g \in L^2(U)$. Then there exists a unique weak solution $u(t)$ of* (5.11) *with $\mathcal{A}^t$ in divergence form. For this solution we have $u \in L^2(0,T;H^1_0(U))$ and $\frac{du}{dt} \in L^2(0,T;H^{-1}(U))$.*

Proof. We begin with the proof of existence. The sequences of Galerkin's approximations $u_m \in L^2(0,T;H^1_0(U))$ and $\frac{d}{dt}u_m \in L^2(0,T;H^{-1}(U))$ are bounded due to the estimates of the previous theorem. Thus there exists a function $u \in$

$L^2(0,T;H_0^1(U))$ with $\frac{du}{dt} \in L^2(0,T;H^{-1}(U))$ and a weakly convergent subsequence (which we also index with m)

$$u_m \rightharpoonup u, \quad \text{in } L^2(0,T;H_0^1(U)),$$
$$\tfrac{d}{dt}u_m \rightharpoonup \tfrac{d}{dt}u, \quad \text{in } L^2(0,T;H^{-1}(U)).$$

Let us define

$$v = \sum_{k=1}^{N} w^k(t)\psi_k, \tag{5.17}$$

with smooth $w^k(t)$ such that $v \in C^1(0,T;H_0^1(U))$. We multiply (5.12) for $m \geq N$ by $w^k(t)$, sum with respect to k from 1 to N, and integrate the sum with respect to t in $[0,T]$

$$\int_0^T \big(\langle \tfrac{d}{dt}u_m(t), v(t)\rangle + B^t[u_m(t), v(t)]\big)dt = \int_0^T (f(t), v(t))dt.$$

Passing to the limit with m (weak convergence) we obtain

$$\int_0^T \big(\langle \tfrac{d}{dt}u(t), v(t)\rangle + B^t[u(t), v(t)]\big)dt = \int_0^T (f(t), v(t))dt, \tag{5.18}$$

where the equality holds for any $v \in L^2(0,T;H_0^1(U))$, as functions of the form (5.17) are dense in $L^2(0,T;H_0^1(U))$. Since $v \in L^2(0,T;H_0^1(U))$ is arbitrary, we obtain for each $z \in H_0^1(U)$ and a.e. $t \in [0,T]$ the equality

$$\langle \tfrac{d}{dt}u(t), z\rangle + B^t[u(t), z] = (f(t), z). \tag{5.19}$$

Furthermore, $u \in C([0,T];L^2(U))$ by Theorem 5.16.

To prove $u(0) = g$ we use the smoothness of $w^k(t)$. Integrating by parts we can write (5.18) as

$$\int_0^T \big(-\langle \tfrac{d}{dt}v(t), u(t)\rangle + B^t[u(t), v(t)]\big)dt = \int_0^T (f(t), v(t))dt + (u(0), v(0))$$

for $v \in C^1(0,T;H_0^1(U))$ such that $v(T) = 0$.

Similarly

$$\int_0^T \big(-\langle \tfrac{d}{dt}v(t), u_m(t)\rangle + B^t[u_m(t), v(t)]\big)dt = \int_0^T (f(t), v(t))dt + (u_m(0), v(0)).$$

Passing in the last equality to the limit with m we obtain

$$\int_0^T \big(-\langle \tfrac{d}{dt}v(t), u(t)\rangle + B^t[u(t), v(t)]\big)dt = \int_0^T (f(t), v(t))dt + (g, v(0)),$$

since $u_m(0) \to g$ in $L^2(U)$, which follows from the initial condition $d_m^k(0) = (g, \psi_k)$ in Theorem 5.26. As $v(0)$ is arbitrary, we get $u(0) = g$.

Let us now prove uniqueness. Choosing $f = 0$, $g = 0$ we will prove $u = 0$.

Inserting $z = u(t)$ in (5.19) we get

$$\frac{d}{dt}\Big(\frac{1}{2}\|u(t)\|^2_{L^2(U)}\Big) + B^t[u(t), u(t)] = 0,$$

for a.e. t.

Since by Theorem 5.20, we have the inequality

$$B^t[u(t), u(t)] \geq \beta\|u(t)\|^2_{H_0^1(U)} - \gamma\|u(t)\|^2_{L^2(U)} \geq -\gamma\|u(t)\|^2_{L^2(U)},$$

then

$$\frac{d}{dt}\Big(\frac{1}{2}\|u(t)\|^2_{L^2(U)}\Big) - \gamma\|u(t)\|^2_{L^2(U)} \leq 0.$$

By Gronwall's lemma we get $u = 0$ for $u(0) = g = 0$. ∎

THEOREM. 5.29 (Regularity of solutions) *Let the coefficients a_{ij}, b_i $(i, j = 1, \ldots, d)$ and c are C^1 in x and do not depend on t. Assume $g \in H_0^1(U)$ and $f \in L^2(0, T; L^2(U))$ with ∂U smooth enough. Then u which is a weak solution of* (5.11) *such that $u \in L^2(0, T; H_0^1(U))$, $\frac{du}{dt} \in L^2(0, T; H^{-1}(U))$ is in fact more regular*

$$u \in L^2(0, T; H^2(U)) \cap L^\infty(0, T; H_0^1(U)), \quad \tfrac{du}{dt} \in L^2(0, T; L^2(U)).$$

If, in addition, $g \in H^2(U)$, $\frac{df}{dt} \in L^2(0, T; L^2(U))$, then

$$u \in L^\infty(0, T; H^2(U)), \quad \tfrac{du}{dt} \in L^\infty(0, T; L^2(U)) \cap L^2(0, T; H_0^1(U)),$$
$$\tfrac{d^2u}{dt^2} \in L^2(0, T; H^{-1}(U)).$$

Remark. 5.3 *For $g \in H^{2m+1}(U) \cap H_0^1(U)$, $\frac{d^k f}{dt^k} \in L^2(0, T; H^{2m-2k}(U))$, $k = 0, 1, \ldots, m$, we can show that $\frac{d^k u}{dt^k} \in L^2(0, T; H^{2m+2-2k}(U))$, $k = 0, 1, \ldots, m + 1$, if ∂U is sufficiently smooth.*

THEOREM. 5.30 (Maximum principle) *Let $\mathcal{A}^t$ be given in the nondivergence form with $c = 0$ and $u \in C^{1,2}(U_T) \cap C(\bar{U}_T)$.*
If

$$\tfrac{\partial}{\partial t}u + \mathcal{A}^t u \leq 0 \quad \text{in } U_T,$$

then

$$\max_{\bar{U}_T} u = \max_{\Gamma_T} u, \quad \text{where } \Gamma_T = \bar{U}_T \setminus U_T.$$

If

$$\frac{\partial}{\partial t}u + \mathcal{A}^t u \geq 0 \quad \textit{in } U_T,$$

then

$$\min_{\bar{U}_T} u = \min_{\Gamma_T} u.$$

Proof. We will present the proof for the strict inequality $\frac{\partial}{\partial t}u + \mathcal{A}^t u < 0$ only. Let $u(t_0, x_0) = \max_{\bar{U}_T} u$, where $(t_0, x_0) \in U_T$. If $0 < t_0 < T$, then (t_0, x_0) belong to the interior of U_T (U is open). Then $\frac{\partial u}{\partial t}(t_0, x_0) = 0$, since this point is a maximum of u.

On the other hand, $\mathcal{A}^t u \geq 0$ at (t_0, x_0). This claim is due to the following observations: at maximum the matrix of second derivatives

$$\left\{ \frac{\partial^2 u}{\partial x_i \partial x_j} \right\}_{i,j=1}^{d}$$

is negative semi-definite, and the matrix of the coefficients

$$\Big\{ a_{ij}(t_0, x_0) \Big\}_{i,j=1}^{d}$$

is symmetric and positive definite ($\mathcal{A}^t$ is uniformly parabolic). Thus there exists an orthogonal transformation $S = \{s^{ij}\}_{i,j=1}^{d}$, which transforms the matrix of the coefficients into a diagonal matrix $\{e_{ii}\}_{i,j=1}^{d}$ with $e_{ii} > 0$ for each i. Performing the change of variables $y = x_0 + S(x - x_0)$ we get

$$\sum_{i,j=1}^{d} a_{ij}(t_0, x_0) \frac{\partial^2 u(t_0, x_0)}{\partial x_i \partial x_j} = \sum_{i=1}^{d} e_{ii} \frac{\partial^2 u(t_0, x_0)}{\partial y_i^2} \leq 0.$$

On the other hand, the vector of the first derivatives

$$\left\{ \frac{\partial u}{\partial x_i} \right\}_{i=1}^{d}$$

is zero (maximum). Hence

$$\mathcal{A}^t u(t_0, x_0) = -\sum_{i,j=1}^{d} a_{ij}(t_0, x_0) \frac{\partial^2 u(t_0, x_0)}{\partial x_i \partial x_j} + \sum_{i=1}^{d} b_i(t_0, x_0) \frac{\partial u(t_0, x_0)}{\partial x_i} \geq 0.$$

Thus $\frac{\partial}{\partial t}u + \mathcal{A}^t u \geq 0$ at (t_0, x_0),which contradicts the assumption $\frac{\partial}{\partial t}u + \mathcal{A}^t u < 0$.

If $t_0 = T$, then as (t_0, x_0) is the point of maximum, we have $\frac{\partial}{\partial t}u \geq 0$ at (t_0, x_0) (u grows as $t \nearrow t_0$ since at t_0 is a maximum). Then $\frac{\partial}{\partial t}u + \mathcal{A}^t u \geq 0$ (the inequality $\mathcal{A}^t u \geq 0$ is still valid), which again gives a contradiction. ■

THEOREM. 5.31 *Let $\mathcal{A}^t$ be given in the nondivergence form with $c \geq 0$ in U_T and $u \in C^{1,2}(U_T) \cap C(\bar{U}_T)$.*
If

$$\tfrac{\partial}{\partial t}u + \mathcal{A}^t u \leq 0 \quad \textit{in } U_T,$$

then

$$\max_{\bar{U}_T} u \leq \max_{\Gamma_T} u^+,$$

where $u^+ = \max(u, 0)$.
If

$$\tfrac{\partial}{\partial t}u + \mathcal{A}^t u \geq 0 \quad \textit{in } U_T,$$

then

$$\min_{\bar{U}_T} u \geq -\max_{\Gamma_T} u^-,$$

where $u^- = \max(-u, 0)$.

COROLLARY. 5.32 (Comparison principle) *Let $u, v \in C^{1,2}(U_T) \cap C(\bar{U}_T)$ and $\mathcal{A}^t$ be given in the nondivergence form with $c \geq 0$.*
If

$$\begin{aligned} &\tfrac{\partial}{\partial t}u + \mathcal{A}^t u \geq 0 \quad \textit{in } U_T, \\ &\tfrac{\partial}{\partial t}v + \mathcal{A}^t v \leq 0 \quad \textit{in } U_T, \\ &u \geq v \quad \textit{on } \Gamma_T, \end{aligned}$$

then $u \geq v$ in U_T.

Proof. Taking $z = u - v$ we obtain

$$\begin{aligned} &\tfrac{\partial}{\partial t}z + \mathcal{A}^t z \geq 0 \quad \text{in } U_T, \\ &z \geq 0 \quad \text{on } \Gamma_T. \end{aligned}$$

By Theorem 5.31 we get $z \geq 0$ in U_T. ■

5.4 The Black-Scholes equation

The Black-Scholes equation is a standard example of a second-order parabolic equation in finance. Let us recall that the equation describes the time dynamics of option prices as functions of the underlying price s and time t

$$\frac{\partial V(t,s)}{\partial t} + \frac{1}{2}\sigma^2 s^2 \frac{\partial^2 V(t,s)}{\partial s^2} + rs\frac{\partial V(t,s)}{\partial s} - rV(t,s) = 0, \tag{5.20}$$

with the terminal condition which is the option payoff

$$V(T,s) = g(s),$$

and the problem is defined on $[0,T] \times [0,\infty)$.

The presented earlier theory of parabolic equations cannot be directly applied to the Black-Scholes equation:

1. The domain is unbounded as $s \in [0,\infty)$.
2. The coefficients in the equation are unbounded.

These problems can be solved by:

1. Building a separate theory suitable for the Black-Scholes equation. Such a theory requires special functional spaces in which we mimic the Sobolev space approach presented above. We will describe briefly that method later on.
2. Making a suitable change of variables to reduce the Black-Scholes equation to one of the problems analyzed in this chapter. We describe that approach below, omitting technical details.

We can eliminate the unbounded coefficients substituting $x = \ln s$. We also substitute $\tau = T - t$ replacing terminal conditions by initial conditions. Due to these substitutions we obtain the initial value problem

$$\begin{aligned} &\frac{\partial u}{\partial \tau} - \frac{1}{2}\tilde{\sigma}^2(\tau,x)\frac{\partial^2 u}{\partial x^2} + b(\tau,x)\frac{\partial u}{\partial x} + c(\tau,x)u = 0, \\ &u(0,x) = \tilde{g}(x), \end{aligned} \tag{5.21}$$

where $u(\tau,x) = V(T-\tau,e^x)$. The coefficients $\tilde{\sigma}$, b, and c are bounded (the boundedness follows by their financial meaning) but the domain is unbounded: $x \in \mathbb{R}$ ($\tilde{\sigma}$, $\tilde{g}$ denote functions σ and g after the above change of variables).

Assuming additionally that $\tilde{\sigma}(\tau,\cdot)$ is of class C^1 in x and the following estimates are uniform in τ

$$0 < \sigma_L \le \tilde{\sigma}(\tau,x) \le \sigma_U, \quad \left|\frac{\partial \tilde{\sigma}(\tau,x)}{\partial x}\right| \le C, \tag{5.22}$$

we can transform (5.21) to the divergence form

$$\begin{aligned} &\frac{\partial u}{\partial \tau} - \frac{1}{2}\frac{\partial}{\partial x}\Big(\tilde{\sigma}^2(\tau,x)\frac{\partial u}{\partial x}\Big) + \bar{b}(\tau,x)\frac{\partial u}{\partial x} + c(\tau,x)u = 0, \\ &u(0,x) = \tilde{g}(x). \end{aligned} \tag{5.23}$$

Equation (5.23) is an example of second-order parabolic equations in an unbounded domain $U_T = (0,T] \times \mathbb{R}^d$. Such equations are investigated in weighted Sobolev spaces. Let us consider the general case

$$\begin{aligned} &\tfrac{\partial}{\partial t}u + \mathcal{A}^t u = f, \quad \text{in } U_T, \\ &u = g, \quad \text{on } \{t=0\} \times \mathbb{R}^d, \end{aligned} \tag{5.24}$$

where $\frac{\partial}{\partial t} + \mathcal{A}^t$ is uniformly parabolic and

$$\mathcal{A}^t u(t,x) = \sum_{i,j=1}^{d} -\frac{\partial}{\partial x_i}\left(a_{ij}(t,x)\frac{\partial u(t,x)}{\partial x_j}\right) + \sum_{i=1}^{d} b_i(t,x)\frac{\partial u(t,x)}{\partial x_i} + c(t,x)u(t,x).$$

Let $L^2_\rho(\mathbb{R}^d)$ denote the space with weight ρ

$$L^2_\rho(\mathbb{R}^d) = \Big\{u\colon \int_{\mathbb{R}^d} |u(x)|^2 \rho(x)dx < +\infty\Big\}.$$

The norm in that space is defined by the formula

$$\|u\|_{L^2_\rho} = \Big(\int_{\mathbb{R}^d} |u(x)|^2 \rho(x)dx\Big)^{1/2}.$$

We define also spaces $H^k_\rho(\mathbb{R}^d)$ with the norm

$$\|u\|_{H^k_\rho} = \|u\|_{L^2_\rho} + \sum_{|\alpha|\le k} \|D^\alpha u\|_{L^2_\rho}.$$

DEFINITION. 5.33 *The weight functions $\rho(x)$ are functions of class $C^1(\mathbb{R}^d)$ such that $\rho(x)^{-1}D\rho(x)$ are bounded uniformly in $x \in \mathbb{R}^d$.*

The proof of existence and uniqueness of problem (5.24) is analogous to the proof for a bounded domain U.

First, integrating by parts the second-order terms in $\mathcal{A}^t$ we define for every $u, v \in H^1_\rho(\mathbb{R}^d)$ the bilinear form

$$B^t_\rho[u,v] = \int_{\mathbb{R}^d}\Big(\sum_{i,j=1}^{d} a_{ij}(t,x)\frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_j} + \sum_{i=1}^{d}\hat{b}_i(t,x)\frac{\partial u}{\partial x_i}v + c(t,x)uv\Big)\rho(x)dx, \tag{5.25}$$

where

$$\hat{b}_i(t,x) = b_i(t,x) + \sum_{j=1}^{d} a_{ij}(t,x)\frac{1}{\rho(x)}\frac{\partial \rho(x)}{\partial x_j}.$$

Assuming a_{ij}, b_i, $c \in L^\infty(U_T)$ $(i,j = 1,\dots,d)$ and the symmetry $a_{ij} = a_{ji}$ we can obtain the energy estimates for $B^t_\rho[u,v]$ with $u, v \in H^1_\rho(\mathbb{R}^d)$

$$|B^t_\rho[u,v]| \le \alpha \|u\|_{H^1_\rho} \|v\|_{H^1_\rho},$$
$$\beta \|u\|^2_{H^1_\rho} \le B^t_\rho[u,u] + \gamma \|u\|^2_{L^2_\rho}.$$

The proof of the above estimates is similar to the proof for a bounded U. The only difference is in the passage from the estimate

$$\beta \|Du\|^2_{L^2_\rho} \le B^t_\rho[u,u] + \gamma \|u\|^2_{L^2_\rho}$$

to the estimate

$$\beta \|u\|^2_{H^1_\rho} \le B^t_\rho[u,u] + \gamma \|u\|^2_{L^2_\rho}.$$

For a bounded U we have used Poincaré's inequality. In $\mathbb{R}^d$ we replace $\|Du\|^2_{L^2_\rho}$ by $\|u\|^2_{H^1_\rho}$ by adding $\beta\|u\|^2_{L^2_\rho}$ to both sides which only increases the constant γ.

Like in the case of a bounded U we look at a solution of (5.24) as a mapping $u\colon\ [0,T] \to H^1_\rho(\mathbb{R}^d)$. Multiplying (5.24) by $\rho(x)v(x)$, with $v \in H^1_\rho(\mathbb{R}^d)$, and integrating on $[0,T]$ we obtain the weak form of the equation

$$\begin{aligned} &\langle \tfrac{d}{dt} u(t), v\rangle + B^t_\rho[u(t),v] = \langle f(t), v\rangle, \\ &u(0) = g, \end{aligned} \tag{5.26}$$

where $\langle v^*, v\rangle$ denotes the value of functionals v^* from H^{-1}_ρ, the space dual to H^1_ρ, on elements $v \in H^1_\rho$. Similarly as for a bounded U, H^1_ρ is densely embedded in L^2_ρ, which from its part is densely embedded in H^{-1}_ρ.

Remark. 5.4 *To prove the existence of solutions for* (5.24) *by the Lax-Milgram theorem we need coerciveness of the bilinear form* $B^t_\rho[u,u]$

$$\beta \|u\|^2_{H^1_\rho} \le B^t_\rho[u,u].$$

This cannot be obtained by adding to $c(x)$ *a large constant like in Theorem 5.21 since now* $c(x)$ *has a clear financial meaning. Then we can achieve the same effect by the change of variables* $u_\gamma(t) = \mathrm{e}^{-\gamma t}u(t)$. *Then* $u_\gamma(t)$ *solves* (5.24) *in which the bilinear form* $B^t_\rho[u,v]$ *is replaced by*

$$B^t_{\rho,\gamma}[u,v] = B^t_\rho[u,v] + \gamma(u,v)_{L^2_\rho},$$

where $(u,v)_{L^2_\rho}$ *denotes the scalar product in* $L^2_\rho(\mathbb{R}^d)$. *It is obvious that* $B^t_{\rho,\gamma}[u,u]$ *is already coercive.*

Making the change of variables described above, integrating equation (5.26) on $[0,\tau]$ and applying the coerciveness of $B^t_{\rho,\gamma}[u,u]$, we obtain the estimate for $\tau \in [0,T]$

$$\mathrm{e}^{-2\gamma\tau}\|u(\tau)\|^2_{L^2_\rho}+(2\beta-\epsilon)\int_0^\tau \mathrm{e}^{-2\gamma s}\|u(s)\|^2_{H^1_\rho}ds \le \|g\|^2_{L^2_\rho}+\epsilon^{-1}\int_0^\tau \|f(s)\|^2_{H^{-1}_\rho}ds, \tag{5.27}$$

where β is the constant from the coerciveness estimate for $B^t_{\rho,\gamma}[u,u]$ and $\epsilon>0$ is a small constant.

Estimate (5.27) is the energy estimate for (5.24). Due to this estimate, the Galerkin approximation of (5.24) has a convergent subsequence. Then we have the following theorem.

THEOREM. 5.34 *Assume a_{ij}, b_i, $c \in L^\infty(U_T)$ $(i,j=1,\dots,d)$, $g \in L^2_\rho$ and $f \in L^2(0,T;H^{-1}_\rho)$. Then there exists a unique weak solution $u(t)$ of the initial value problem* (5.24) *such that $u \in L^2(0,T;H^1_\rho)$ with $\frac{du}{dt} \in L^2(0,T;H^{-1}_\rho)$ and the following estimate holds for $\tau \in [0,T]$*

$$\mathrm{e}^{-2\gamma\tau}\|u(\tau)\|^2_{L^2_\rho}+(2\beta-\epsilon)\int_0^\tau \mathrm{e}^{-2\gamma s}\|u(s)\|^2_{H^1_\rho}ds \le \|g\|^2_{L^2_\rho}+\epsilon^{-1}\int_0^\tau \|f(s)\|^2_{H^{-1}_\rho}ds.$$

The above theorem enables us to make the right choice of the weight function $\rho(x)$. For the theorem to work we need $g \in L^2_\rho$. As g is the option payoff, we have to choose ρ that makes this payoff an element of L^2_ρ.

Consider as an example a vanilla call option with payoff $(s-K)^+$. Passing to $x=\ln s$ we get $g(x)=\left(\mathrm{e}^x-K\right)^+$. Then a proper weight function is $\rho(x)=\mathrm{e}^{-\lambda\varphi(x)}$, where $\varphi(x)=(1+|x|^2)^{1/2}$ and $\lambda>0$ is sufficiently large.

Remark. 5.5 *Assume that σ and r in the Black-Scholes equation* (5.20) *are constant. Then there is a change of variables which makes the equation particularly simple. Namely, making the following change to the independent variables*

$$x=\ln s, \quad \tau=\frac{\sigma^2}{2}(T-t)$$

and of the dependent variable

$$u(\tau,x):=\exp\left(\frac{1}{2}(q-1)x+\frac{1}{4}(q+1)^2\tau\right)V\left(T-\frac{2\tau}{\sigma^2},\, e^x\right),$$

where $q = \frac{2r}{\sigma^2}$, we obtain the heat equation which is particularly suitable for numerical computations

$$\frac{\partial u(\tau,x)}{\partial \tau} - \frac{\partial^2 u(\tau,x)}{\partial x^2} = 0, \quad \tau \in \left[0, \frac{\sigma^2}{2}T\right], x \in \mathbb{R},$$
$$u(0,x) = e^{\frac{1}{2}(q-1)x} g(e^x), \quad x \in \mathbb{R}.$$

For numerical practice, the existence theorem in $\mathbb{R}^d$ is not so important as obtaining numerical solutions on $\mathbb{R}^d$ is impossible. For numerical computations we have to limit considerations to a bounded domain $U \subset \mathbb{R}^d$ and introduce *artificial* boundaries and *artificial* boundary conditions. For initial-boundary value problems localized on bounded sets, the theory developed in Sobolev spaces without weights is applicable. The choice of a bounded domain U has to be a compromise between the numerical error, which usually decreases with the size of U, and the computational time, which increases with the size of U.

A crucial problem for controlling numerical error is the choice of artificial boundary conditions. In quantitative finance, we can find hints by analyzing the economic aspects of the problem. The nature of such hints is well visible in an example of vanilla options (we consider these options in variables (t,s) leaving to the reader the task of translating the obtained artificial boundary conditions into the variables used in (5.24)). For a call option we have $\lim_{s\to 0} V(t,s) = 0$ which has an obvious economic interpretation. For large s we can approximate the option price by $V(t,s) \approx s(t) - Ke^{-r(T-t)}$. That approximation is obtained by the computation of the present value at t of the payoff $(s(T) - K)$ (valid for large s). For a put option we have $\lim_{s\to 0} V(t,s) = Ke^{-r(T-t)}$, discounted to t payoff K at T, and the obvious condition $\lim_{s\to\infty} V(t,s) = 0$.

Chapter 6

Finite difference methods for parabolic equations

6.1 Introduction to finite differences

Historically, the finite difference method is the first and simplest method for discretizing partial differential equations. We begin the presentation of this method with the model problem of the one-dimensional heat conduction

$$\frac{\partial u(t,x)}{\partial t} - \frac{\partial^2 u(t,x)}{\partial x^2} = 0.$$

We will consider the problem in a bounded interval (x_{min}, x_{max}) for $t \in [0,T]$. We divide the interval (x_{min}, x_{max}) into M subintervals and the time interval $[0,T]$ into N subintervals and set

$$\delta x = \frac{x_{max} - x_{min}}{M}, \quad \delta t = \frac{T}{N}.$$

The points

$$x_k := x_{min} + k \cdot \delta x,\ k = 0, 1, \ldots, M, \quad t_n := n \cdot \delta t,\ n = 0, 1, \ldots, N,$$

are called the *grid points* (or the *mesh points*).

We approximate derivatives by finite differences

$$\frac{\partial u}{\partial t}(t_n, x_k) \approx \frac{u(t_{n+1}, x_k) - u(t_n, x_k)}{\delta t}$$
$$\frac{\partial^2 u}{\partial x^2}(t_n, x_k) \approx \frac{u(t_n, x_{k+1}) - 2u(t_n, x_k) + u(t_n, x_{k-1})}{(\delta x)^2}$$

and denote by w_k^n the approximation of $u(t_n, x_k)$.

We now define several simple difference schemes approximating a solution of the heat equation.

Euler explicit scheme. In the Euler explicit scheme, we approximate the time derivative by a forward difference; this gives

$$\frac{w_k^{n+1} - w_k^n}{\delta t} - \frac{w_{k+1}^n - 2w_k^n + w_{k-1}^n}{(\delta x)^2} = 0.$$

Thus, the approximation of the heat equation satisfies

$$w_k^{n+1} = \lambda w_{k+1}^n + (1 - 2\lambda) w_k^n + \lambda w_{k-1}^n,$$

where $\lambda = \frac{\delta t}{(\delta x)^2}$.

Introducing the vector

$$W^n = (w_1^n, w_2^n, \ldots, w_{M-1}^n),$$

(w_0^n and w_M^n are omitted as known from the boundary conditions) we can write the scheme in the matrix form

$$BW^{n+1} = AW^n + d^n,$$

where B is an identity matrix of dimension $(M-1) \times (M-1)$ and

$$A = \begin{pmatrix} 1-2\lambda & \lambda & & 0 \\ \lambda & \ddots & \lambda & \\ & \ddots & \ddots & \ddots \\ 0 & & \ddots & \ddots \end{pmatrix}, \qquad d^n = \begin{pmatrix} \lambda w_0^n \\ 0 \\ \vdots \\ 0 \\ \lambda w_M^n \end{pmatrix},$$

with A of dimension $(M-1) \times (M-1)$ and d^n of dimension $M-1$.

Euler implicit scheme. In that scheme, we approximate the time derivative by a backward difference to obtain

$$\frac{w_k^{n+1} - w_k^n}{\delta t} - \frac{w_{k+1}^{n+1} - 2w_k^{n+1} + w_{k-1}^{n+1}}{(\delta x)^2} = 0.$$

Then we get the equation

$$-\lambda w_{k+1}^{n+1} + (1 + 2\lambda) w_k^{n+1} - \lambda w_{k-1}^{n+1} = w_k^n,$$

which can be written in the matrix form

$$BW^{n+1} = AW^n + d^n,$$

where A is an identity matrix of dimension $(M-1)\times(M-1)$ and

$$B = \begin{pmatrix} 1+2\lambda & -\lambda & & 0 \\ -\lambda & \ddots & -\lambda & \\ & \ddots & \ddots & \ddots \\ 0 & & \ddots & \ddots \end{pmatrix}, \qquad d^n = \begin{pmatrix} \lambda w_0^{n+1} \\ 0 \\ \vdots \\ 0 \\ \lambda w_M^{n+1} \end{pmatrix},$$

where B is $(M-1)\times(M-1)$-dimensional and d^n, $(M-1)$-dimensional.

Crank-Nicolson scheme. That scheme can be considered as a "linear combination" of the explicit and implicit Euler schemes

$$-\frac{\lambda}{2}w_{k-1}^{n+1} + (1+\lambda)w_k^{n+1} - \frac{\lambda}{2}w_{k+1}^{n+1} = \frac{\lambda}{2}w_{k-1}^n + (1-\lambda)w_k^n + \frac{\lambda}{2}w_{k+1}^n.$$

The matrix form of that scheme looks similar to the former two schemes

$$BW^{n+1} = AW^n + d^n,$$

but with different matrices

$$B = \begin{pmatrix} 1+\lambda & -\frac{\lambda}{2} & & 0 \\ -\frac{\lambda}{2} & \ddots & -\frac{\lambda}{2} & \\ & \ddots & \ddots & \ddots \\ 0 & & \ddots & \ddots \end{pmatrix}, \qquad d^n = \frac{\lambda}{2}\begin{pmatrix} w_0^n + w_0^{n+1} \\ 0 \\ \vdots \\ 0 \\ w_M^n + w_M^{n+1} \end{pmatrix},$$

$$A = \begin{pmatrix} 1-\lambda & \frac{\lambda}{2} & & 0 \\ \frac{\lambda}{2} & \ddots & \frac{\lambda}{2} & \\ & \ddots & \ddots & \ddots \\ 0 & & \ddots & \ddots \end{pmatrix}.$$

Consider now a generalized one-dimensional parabolic equation (the Black-Scholes equation (5.21) is of this form)

$$\frac{\partial u}{\partial t} - a^2(t,x)\frac{\partial^2 u}{\partial x^2} + b(t,x)\frac{\partial u}{\partial x} + c(t,x)u = 0. \tag{6.1}$$

To obtain a finite difference approximation of this equation, we approximate the first-order x derivative by the expression

$$\frac{\partial u}{\partial x}(t_n, x_k) \approx \frac{u(t_n, x_{k+1}) - u(t_n, x_{k-1})}{2\delta x}.$$

Denoting, as previously, by w_k^n the approximation of $u(t_n, x_k)$ we obtain for the schemes defined above the matrix equation

$$B^{n+1}W^{n+1} = A^n W^n + d^n.$$

Since the coefficients in (6.1) are time-dependent, we obtain different matrices A^n and B^n on different time levels. The forms of A^n, B^{n+1} and d^n depend on the scheme used.

For the Euler explicit scheme B^{n+1} is an identity matrix and A^n is tridiagonal with the elements:

$$\begin{aligned}
A^n_{k,k} &= 1 - 2\lambda\big(a^n_k\big)^2 - \delta t\, c^n_k,\\
A^n_{k,k+1} &= -\frac{\gamma}{2} b^n_{k+1} + \lambda\big(a^n_{k+1}\big)^2,\\
A^n_{k,k-1} &= \frac{\gamma}{2} b^n_{k-1} + \lambda\big(a^n_{k-1}\big)^2,
\end{aligned}$$

where

$$\lambda = \frac{\delta t}{(\delta x)^2}, \quad \gamma = \frac{\delta t}{\delta x}.$$

The vector d^n has all zero elements but the first and last

$$d^n_1 = \lambda\big(a^n_0\big)^2 w^n_0 + \frac{\gamma}{2} b^n_0 w^n_0, \quad d^n_{M-1} = \lambda\big(a^n_M\big)^2 w^n_M - \frac{\gamma}{2} b^n_M w^n_M.$$

For the Euler implicit scheme A^n is an identity matrix and B^{n+1} is tridiagonal with the elements:

$$\begin{aligned}
B^{n+1}_{k,k} &= 1 + 2\lambda\big(a^{n+1}_k\big)^2 + \delta t\, c^n_k,\\
B^{n+1}_{k,k+1} &= \frac{\gamma}{2} b^{n+1}_{k+1} - \lambda\big(a^{n+1}_{k+1}\big)^2,\\
B^{n+1}_{k,k-1} &= -\frac{\gamma}{2} b^{n+1}_{k-1} - \lambda\big(a^{n+1}_{k-1}\big)^2.
\end{aligned}$$

For the first and the last elements of d^n we have

$$\begin{aligned}
d^n_1 &= \lambda\big(a^{n+1}_0\big)^2 w^{n+1}_0 + \frac{\gamma}{2} b^{n+1}_0 w^{n+1}_0,\\
d^n_{M-1} &= \lambda\big(a^{n+1}_M\big)^2 w^{n+1}_M - \frac{\gamma}{2} b^{n+1}_M w^{n+1}_M.
\end{aligned}$$

For the Crank-Nicolson scheme the corresponding matrices are tridiagonal with the elements:

$$
\begin{aligned}
B^{n+1}_{k,k} &= 1 + \lambda\big(a^{n+1}_k\big)^2 + \frac{\delta t}{2} c^n_k,\\
B^{n+1}_{k,k+1} &= \frac{\gamma}{4} b^{n+1}_{k+1} - \frac{\lambda}{2}\big(a^{n+1}_{k+1}\big)^2,\\
B^{n+1}_{k,k-1} &= -\frac{\gamma}{4} b^{n+1}_{k-1} - \frac{\lambda}{2}\big(a^{n+1}_{k-1}\big)^2.
\end{aligned}
$$

$$
\begin{aligned}
A^{n}_{k,k} &= 1 - \lambda\big(a^{n}_k\big)^2 - \frac{\delta t}{2} c^n_k,\\
A^{n}_{k,k+1} &= -\frac{\gamma}{4} b^{n}_{k+1} + \frac{\lambda}{2}\big(a^{n}_{k+1}\big)^2,\\
A^{n}_{k,k-1} &= \frac{\gamma}{4} b^{n}_{k-1} + \frac{\lambda}{2}\big(a^{n}_{k-1}\big)^2.
\end{aligned}
$$

For d^n we have

$$
\begin{aligned}
d^n_1 &= \frac{\lambda}{2}\Big(\big(a^n_0\big)^2 w^n_0 + \big(a^{n+1}_0\big)^2 w^{n+1}_0\Big) + \frac{\gamma}{4}\Big(b^n_0 w^n_0 + b^{n+1}_0 w^{n+1}_0\Big),\\
d^n_{M-1} &= \frac{\lambda}{2}\Big(\big(a^n_M\big)^2 w^n_M + \big(a^{n+1}_M\big)^2 w^{n+1}_M\Big) - \frac{\gamma}{4}\Big(b^n_M w^n_M + b^{n+1}_M w^{n+1}_M\Big).
\end{aligned}
$$

6.2 Convergence analysis of two level schemes

We will restrict the analysis of convergence to two-time-level schemes for the linear parabolic initial-boundary value problem

$$
\begin{aligned}
&\tfrac{\partial}{\partial t} u + \mathcal{A}^t u = f, \quad \text{in } (0,T]\times U,\\
&u = 0, \quad \text{on } [0,T]\times \partial U,\\
&u = g, \quad \text{on } \{t=0\}\times U,
\end{aligned}
\tag{6.2}
$$

where $\mathcal{A}^t$ is the uniformly parabolic operator of Definition 5.24.

We introduce a grid in U i.e. a discrete set J_U of grid points $x_{\boldsymbol{k}}$ indexed by the vectors $\boldsymbol{k}$ with components k_i, $i = 1,\dots,d$. For $U \subset \mathbb{R}^d$ the discretization spacing δx_i can depend on the direction x_i but all δx_i are of the same order h and we assume that all $\delta x_i \to 0$ with $h \to 0$. The interval $[0,T]$ is discretized with points t_n, $n = 0,1,\dots N$ with the time step δt (this time step can be a function of h as well) such that $N\delta t = T$.

Let us define the localization operator

$$I_{\delta t,h}\colon\ C([0,T]\times U)\to\mathbb{R}^{(N+1)\times\#J_U},$$

which maps continuous functions to their values at the grid nodes $w^n_{\boldsymbol{k}}=w(t_n,x_{\boldsymbol{k}})$, $x_{\boldsymbol{k}}\in J_U$, $n=0,1,\dots,N$. For each t_n we introduce a grid function $W^n=\{w^n_{\boldsymbol{k}}\}$ which is defined in the grid points of J_U.

We approximate derivatives in the differential operator $\mathcal{A}^t$ by finite differences in the x variable and the time derivative by a forward or backward finite difference. Then we obtain the following finite difference approximation for the differential problem (6.2)

$$L^{n+1}W^{n+1}=R^nW^n+F^n,$$

where the terms for time level $n+1$ are on the left hand side of the equation and the terms for time level n are on the right hand side, and F^n is the localization of the nonhomogeneous term f. For the one-dimensional schemes of the previous section, matrices L^n and R^n straightforwardly correspond to matrices B^n and A^n

$$B^n=\delta tL^n,\quad A^n=\delta tR^n.$$

DEFINITION. 6.1 *The system of equations*

$$L^{n+1}W^{n+1}=R^nW^n+F^n,\quad n=0,1,\dots,N-1, \tag{6.3}$$

is called the two-time-level difference scheme for problem (6.2).

The difference scheme (6.3) *is said to be* solvable *if*

$$\|(L^n)^{-1}\|\le C\delta t,$$

where the constant C is independent of n.

The solvability condition means that starting from W^0 we can compute the grid function W^n for all subsequent time levels.

If u is a smooth function and W^n is a grid function obtained by the localization of u, then we can expect the convergence $L^{n+1}W^{n+1}-R^nW^n-F^n\to\frac{\partial u}{\partial t}+\mathcal{A}^tu-f$, as $\delta t,h\to 0$. That leads to the notion of *consistency* of a difference scheme.

DEFINITION. 6.2 *For sufficiently smooth functions φ we define the truncation error*

$$\Psi^n(\varphi)=L^{n+1}(I_{\delta t,h}\varphi)^{n+1}-R^n(I_{\delta t,h}\varphi)^n-F^n-\big(I_{\delta t,h}(\tfrac{\partial\varphi}{\partial t}+\mathcal{A}^t\varphi-f)\big)^n,$$

where $(I_{\delta t,h}\varphi)^n$ denotes the localization of φ in the n-th time level.

It is said that the difference scheme (6.3) *is* consistent *with the differential problem* (6.2) *in the norm* $\|\cdot\|$*, if*

$$\lim_{h,\delta t\to 0}\|\Psi^n(\varphi)\|\to 0 \quad \textit{for } n=0,\dots,N-1.$$

When there exist constants $C(\varphi)$*,* k_1*,* k_2 *such that*

$$\|\Psi^n(\varphi)\|\le C(\varphi)\big(|h|^{k_1}+|\delta t|^{k_2}\big),$$

the scheme is said to have the order of approximation k_1 *in* x *and* k_2 *in* t*.*

If the initial-boundary value problem (6.2) *possesses a smooth solution then, to simplify the definition, it is often assumed that* φ *is just this solution. Then the truncation error is*

$$\Psi^n(\varphi)=L^{n+1}(I_{\delta t,h}\varphi)^{n+1}-R^n(I_{\delta t,h}\varphi)^n-F^n.$$

DEFINITION. 6.3 *The difference scheme* (6.3) *is said to be* stable *with respect to initial conditions in the norm* $\|\cdot\|$*, if for any two solutions* W^n*,* V^n *with the initial conditions* W^0*,* V^0*, respectively, but with the same right hand side* F^n*, there exists a constant* C *such that*

$$\|W^n-V^n\|\le C\|W^0-V^0\|,\quad n\delta t\le T.$$

If R^n *and* L^n *are time independent (they are constant in* n*) then the stability condition for the two-time-level scheme* (6.3) *can be written as*

$$\|(L^{-1}R)^n\|\le C,\quad n\delta t\le T.$$

It is possible to investigate the stability with respect to the nonhomogeneous term F^n, but that analysis will be not carried on.

The above defined stability of a difference scheme is in a close relationship to the well-posedness of the corresponding differential problem.

DEFINITION. 6.4 *The problem* (6.2) *is called* well-posed *in the norm* $\|\cdot\|$ *if*

1. *There exists a unique solution of problem* (6.2) *for each initial data* g *such that* $\|g\|<\infty$*.*

2. *There is a constant* C *such that for any two solutions* u_1*,* u_2 *of problem* (6.2) *with data* g_1*,* g_2*,* f_1*,* f_2*, respectively, we have the estimate*

$$\|u_1(t)-u_2(t)\|\le C\big(\|g_1-g_2\|+\sup_{\tau\in[0,t]}\|f_1(\tau)-f_2(\tau)\|\big)\ \textit{for } t\le T.$$

The following theorem explains the connection between consistency, stability, and convergence of difference schemes for well-posed differential problems.

THEOREM. 6.5 (Lax equivalence theorem) *If the problem* (6.2) *is linear, well-posed, and its discrete approximation is solvable and consistent, then a solution of the difference scheme* (6.3) *converges to a solution of* (6.2) *if and only if the difference scheme* (6.3) *is stable.*

Thus, if W^n is a solution of (6.3) *for a given grid J_U and a time discretization with step δt and u is a solution of* (6.2)*, then*

$$\|W^n - (I_{\delta t,h}u)^n\| \to 0 \quad \textit{for } h, \delta t \to 0,\ n\delta t \to t,\ t \le T.$$

When the scheme has the order of approximation k_1 in x and k_2 in t, then there is a constant C such that

$$\|W^n - (I_{\delta t,h}u)^n\| \le C\big(|h|^{k_1} + |\delta t|^{k_2}\big).$$

Proof. W are going to prove only that consistency and stability imply convergence. We assume that the operators L^n and R^n are time independent. Hence, these operators are constant in n and this superscript will be omitted. To simplify notation we write u^n instead of $(I_{\delta t,h}u)^n$.

Assume that u is a smooth solution of (6.2) and consider the difference $W^n - u^n$. Then we have

$$L(W^{n+1} - u^{n+1}) = R(W^n - u^n) - \Psi^n,$$

where Ψ^n is the truncation error. Thus, we obtain

$$W^{n+1} - u^{n+1} = L^{-1}R(W^n - u^n) - L^{-1}\Psi^n,$$

and iterating

$$W^{n+1} - u^{n+1} = -\Big(L^{-1}\Psi^n + L^{-1}RL^{-1}\Psi^{n-1} + \cdots + \big(L^{-1}R\big)^n L^{-1}\Psi^0\Big).$$

By the stability and solvability conditions, we have the inequality

$$\|\big(L^{-1}R\big)^k L^{-1}\| \le C\delta t,$$

which gives the estimate

$$\|W^n - u^n\| \le C\delta t \sum_{i=0}^{n-1} \|\Psi^i\|.$$

If the scheme is consistent then

$$\delta t \sum_{i=0}^{n-1} \|\Psi^i\| \le N\delta t \max_{0\le i\le N-1} \|\Psi^i\| \to 0 \quad \text{as } h, \delta t \to 0,$$

which completes the proof.

Assume now that u is not smooth. For smooth initial values $\tilde{g}$ and nonhomogeneous terms $\tilde{f}$ we obtain smooth solutions of (6.2) by Theorem 5.29 and Remark 5.3. Let $\tilde{g} \in C^\infty(U)$ and $\tilde{f} \in C^\infty(U_T)$ be such that ($\tilde{g}$ and $\tilde{f}$ exist due to Theorem 5.6)

$$\|g - \tilde{g}\| + \sup_{t\in[0,T]} \|f(t) - \tilde{f}(t)\| \le \epsilon.$$

Let now $\tilde{u}$ be a smooth solution of (6.2) with the above $\tilde{g}$ and $\tilde{f}$. By the well-posedness of (6.2), we get

$$\|u(t) - \tilde{u}(t)\| \le C(t)\Big(\|g - \tilde{g}\| + \sup_{\tau\in[0,t]} \|f(\tau) - \tilde{f}(\tau)\|\Big) \le \epsilon C.$$

Taking $\widetilde{W}^n$ a solution of the difference scheme (6.3) corresponding to $\tilde{u}$, we obtain the convergence

$$\begin{aligned}\|W^n - u^n\| &= \|W^n - \widetilde{W}^n + \widetilde{W}^n - \tilde{u}^n + \tilde{u}^n - u^n\| \\ &\le \|W^n - \widetilde{W}^n\| + \|\widetilde{W}^n - \tilde{u}^n\| + \|\tilde{u}^n - u^n\| \to 0, \text{ for } h, \delta t \to 0,\end{aligned}$$

where $\|W^n - \widetilde{W}^n\| \to 0$ follows from the stability of (6.3), $\|\widetilde{W}^n - \tilde{u}^n\| \to 0$, from the consistency, and $\|\tilde{u}^n - u^n\| \to 0$ from the well-posedness of (6.2). ■

6.3 θ-schemes

Let us consider a one-dimensional version of problem (6.2). Since now U is an open interval, without loss of generality we can assume $U = (0, 1)$. Then (6.2) is replaced by

$$\begin{aligned}&\tfrac{\partial}{\partial t}u + \mathcal{A}^t u = f, \quad (t, x) \in (0, T] \times (0, 1),\\ &u(t, 0) = u(t, 1) = 0, \quad t \in [0, T],\\ &u(0, x) = g(x), \quad x \in (0, 1).\end{aligned} \tag{6.4}$$

Assume that the coefficients of $\mathcal{A}^t$ and the nonhomogeneous term f in (6.4) are t independent. Then we can drop the superscript t from $\mathcal{A}^t$ and write

$$\mathcal{A}u = -a^2(x)\tfrac{\partial^2 u}{\partial x^2} + b(x)\tfrac{\partial u}{\partial x} + c(x)u.$$

Thus the Euler explicit and implicit schemes and the Crank-Nicolson scheme can be written by a single formula, the so-called θ-scheme

$$(I+\theta A)W^{n+1} = \big(I-(1-\theta)A\big)W^n + \phi, \tag{6.5}$$

where

$$A_{k,k} = 2\lambda a_k^2 + \delta t\, c_k, \quad A_{k,k+1} = \frac{\gamma}{2}b_{k+1} - \lambda a_{k+1}^2, \quad A_{k,k-1} = -\frac{\gamma}{2}b_{k-1} - \lambda a_{k-1}^2,$$

when, as previously, we denote

$$\lambda = \frac{\delta t}{(\delta x)^2}, \quad \gamma = \frac{\delta t}{\delta x},$$

and ϕ is a localization of f.

To estimate the order of approximation for the θ-schemes we will use the following lemma.

LEMMA. 6.6 *Let $z \in C^4(J)$, $J \subset\subset \mathbb{R}$. Then*

$$\begin{aligned}
&\frac{z(x+h)-2z(x)+z(x-h)}{h^2} = z''(x) + \frac{h^2}{12}z^{(4)}(x) + O(h^3),\\
&\frac{z(x+h)-z(x-h)}{2h} = z'(x) + \frac{h^2}{6}z^{(3)}(x) + O(h^4),\\
&\frac{z(x+h)-z(x)}{h} = z'(x) + \frac{h}{2}z''(x) + O(h^2).
\end{aligned}$$

The proof of this lemma follows easily from the Taylor expansion.

THEOREM. 6.7 *The θ-scheme for the one-dimensional, linear, parabolic problem* (6.4) *has the order of approximation*

$$O\big((\theta-\tfrac{1}{2})\delta t\big) + O\big(|\delta x|^2 + |\delta t|^2\big).$$

Proof. We prove the theorem under the assumption of constant coefficients.

Let y_k^n be a function on the grid, i.e., $y_k^n = y(t_n, x_k)$. Let y denote the vector $\{y_k^n\}_{k=0}^M$ and $\hat{y}$, the vector $\{y_k^{n+1}\}_{k=0}^M$. We will also use the abbreviation $y_\Delta = \frac{\hat{y}-y}{\delta t}$.

Introduce the grid difference operator

$$\Lambda y = -a^2\frac{S^+y - 2y + S^-y}{(\delta x)^2} + b\frac{S^+y - S^-y}{2\delta x} + cy,$$

where S^+ and S^- denote shift operators $S^{\pm}\{y_k^n\} = \{y_{k\pm1}^n\}$. Then the θ-scheme for (6.4) has the form

$$y_\Delta + \Lambda\big(\theta\hat{y} + (1-\theta)y\big) = \phi,$$

where ϕ is a localization of f.

Consider the solution u of (6.4) which is of class $C^{2,4}\big((0,T)\times(0,1)\big)$. Since we are going to consider the action of a finite difference operator on functions of continuous variables, we introduce the following notation:

$$\begin{aligned}
&u = u(\tau,\xi), \quad \text{the localization of } u \text{ at a fixed point } (\tau,\xi),\\
&u_t = \frac{\partial u}{\partial t}(\tau,\xi), \quad u_x = \frac{\partial u}{\partial x}(\tau,\xi),\\
&\hat{u} = u(\tau+\delta t,\xi), \quad \bar{u} = u(\tau+\tfrac{1}{2}\delta t,\xi), \quad u_\Delta = \frac{\hat{u}-u}{\delta t}.
\end{aligned}$$

Since u is a solution of (6.4) then by Definition 6.2 the truncation error ψ is given by the expression

$$\psi = u_\Delta + \Lambda\big(\theta\hat{u} + (1-\theta)u\big) - \phi,$$

and is only a function of u. By Lemma 6.6 we get

$$\begin{aligned}
\Lambda u = -a^2u_{xx} - a^2\frac{(\delta x)^2}{12}u_{xxxx} + O((\delta x)^3) + bu_x + b\frac{(\delta x)^2}{6}u_{xxx} + O((\delta x)^4)\\
+ cu = \mathcal{A}u + \frac{(\delta x)^2}{12}\mathcal{A}u_{xx} + O((\delta x)^2).
\end{aligned}$$

To estimate the truncation error let us note the identities which follow from the definition of u_Δ

$$\begin{aligned}
\hat{u} &= \frac{1}{2}(\hat{u}+u) + \frac{1}{2}(\hat{u}-u) = \frac{1}{2}(\hat{u}+u) + \frac{1}{2}\delta t\, u_\Delta,\\
u &= \frac{1}{2}(\hat{u}+u) - \frac{1}{2}(\hat{u}-u) = \frac{1}{2}(\hat{u}+u) - \frac{1}{2}\delta t\, u_\Delta,\\
\theta\hat{u} + (1-\theta)u &= \frac{1}{2}(\hat{u}+u) + (\theta-1/2)\delta t\, u_\Delta.
\end{aligned}$$

Thus

$$\psi = u_\Delta + \frac{1}{2}\Lambda(\hat{u}+u) + (\theta-1/2)\delta t\Lambda u_\Delta - \phi. \tag{6.6}$$

By Taylor's expansion we get

$$\hat{u} = \bar{u} + \frac{1}{2}\delta t\, \bar{u}_t + \frac{(\delta t)^2}{8}\bar{u}_{tt} + O((\delta t)^3),$$
$$u = \bar{u} - \frac{1}{2}\delta t\, \bar{u}_t + \frac{(\delta t)^2}{8}\bar{u}_{tt} + O((\delta t)^3),$$
$$\frac{1}{2}(\hat{u} + u) = \bar{u} + \frac{(\delta t)^2}{8}\bar{u}_{tt} + O((\delta t)^3).$$

Then comparing

$$u = \frac{1}{2}(\hat{u} + u) - \frac{1}{2}\delta t\, u_\Delta = \bar{u} + \frac{(\delta t)^2}{8}\bar{u}_{tt} - \frac{1}{2}\delta t\, u_\Delta + O((\delta t)^3)$$

with

$$u = \bar{u} - \frac{1}{2}\delta t\, \bar{u}_t + \frac{(\delta t)^2}{8}\bar{u}_{tt} + O((\delta t)^3)$$

we obtain

$$u_\Delta = \bar{u}_t + O((\delta t)^2).$$

Inserting into (6.6) the above expression for u_Δ and the expression for $\frac{1}{2}(\hat{u} + u)$ gives

$$\psi = \bar{u}_t + \mathcal{A}\bar{u} + \frac{(\delta x)^2}{12}\mathcal{A}\bar{u}_{xx} + (\theta - 1/2)\delta t\mathcal{A}\bar{u}_t$$
$$+ \frac{(\delta x)^2}{12}(\theta - 1/2)\delta t\mathcal{A}\bar{u}_{txx} + O((\delta x)^2 + (\delta t)^2) - \phi.$$

We have $\bar{u}_t + \mathcal{A}\bar{u} = \bar{f}$ due to (6.4). Assuming the equality $\bar{f} = \phi$ (f is independent of t and the localization of f is correctly chosen) we have

$$\psi = (\theta - 1/2)\delta t\mathcal{A}\bar{u}_t + O((\delta x)^2 + (\delta t)^2).$$

■

6.4 Stability of difference schemes

We will now analyze the stability of two-time-level difference schemes for the linear, parabolic problem (6.2). To simplify the presentation we assume that the coefficients of (6.2) are time independent and restrict our analysis to the stability with respect to initial data. Let U^n and V^n be two solutions of scheme (6.3) with

initial conditions U^0 and V^0, respectively. Put $W^n = U^n - V^n$. To prove that a scheme is stable in the norm $\|\cdot\|$ we have to prove the following implication

$$LW^{n+1} = RW^n \implies \|W^n\| \le C\|W^0\|, \quad n\delta t \le T,$$

where W^n fulfills zero boundary conditions.

The constant C is independent of n but generally depends on T and with increasing T can lead to an exponential growth of the numerical solution. Only when $C \le 1$ there is no growth, and the solution is bounded by a universal constant for any T.

For parabolic problems, due to a maximum principle, it is very natural to prove stability in the l^∞ norm. The following theorem is a simple example.

THEOREM. 6.8 *The θ-schemes for the heat equation ($a^2 = 1$, $b = 0$, $c = 0$, $f = 0$) with $0 \le \theta \le 1$ and $\theta \ge 1 - \frac{1}{2\lambda}$ yield solutions satisfying*

$$|W^{n+1}|_\infty \le |W^n|_\infty,$$

where $|W^n|_\infty = \max_{0\le k\le M} |w_k^n|$ is the l^∞ norm in $\mathbb{R}^{M+1}$.

That proves the stability

$$|W^n|_\infty \le |W^0|_\infty$$

with the constant $C = 1$.

Proof. The θ-scheme for the heat equation reads

$$\frac{w_k^{n+1} - w_k^n}{\delta t} = \theta \frac{w_{k+1}^{n+1} - 2w_k^{n+1} + w_{k-1}^{n+1}}{(\delta x)^2} + (1-\theta)\frac{w_{k+1}^n - 2w_k^n + w_{k-1}^n}{(\delta x)^2}.$$

We rewrite this equation in the form

$$\begin{aligned}(1 + 2\theta\lambda)w_k^{n+1} =& \theta\lambda(w_{k-1}^{n+1} + w_{k+1}^{n+1}) + (1-\theta)\lambda(w_{k-1}^n + w_{k+1}^n) \\ &+ \big(1 - 2(1-\theta)\lambda\big)w_k^n.\end{aligned}$$

Under the hypothesis of the theorem, all the coefficients on the right hand side are nonnegative. Hence

$$(1+2\theta\lambda)|W^{n+1}|_\infty \le 2\theta\lambda|W^{n+1}|_\infty + 2(1-\theta)\lambda|W^n|_\infty + \big(1-2(1-\theta)\lambda\big)|W^n|_\infty.$$

The rearrangement of terms in the inequality completes the proof. ∎

A maximum principle for parabolic equations can help establish stability in the supremum norm. But there are numerical schemes that do not satisfy a discrete

maximum principle. It appears that a more convenient is the stability analysis in the l^2 norm. In addition, in the l^2 setting, we can use Fourier's analysis which is particularly suitable for constant-coefficient problems. Before passing to the stability analysis of two-time-level schemes let us recall some facts from linear algebra.

LEMMA. 6.9 *Let A be a symmetric matrix in $\mathbb{R}^M$. Then*

$$\rho(A) = \|A\|,$$

where $\|A\|$ denotes the operator norm of A implied by the Euclidean norm in $\mathbb{R}^M$ and $\rho(A)$ denotes the spectral radius of A

$$\rho(A) = \max_{\mu \in \sigma(A)} |\mu|,$$

where $\sigma(A)$ is the spectrum of A, the set of all eigenvalues of A.

For nonsymmetric matrices in $\mathbb{R}^M$, we have only the inequality

$$\rho(A) \le \|A\|.$$

LEMMA. 6.10 *Let A be the matrix defined for the θ-scheme* (6.5) *implied by the operator $\mathcal{A}$ with constant coefficients. The components of A are*

$$A_{k,k} = 2\lambda a^2 + \delta t\, c, \quad A_{k,k+1} = \frac{\gamma}{2} b - \lambda a^2, \quad A_{k,k-1} = -\frac{\gamma}{2} b - \lambda a^2.$$

Then for $\gamma b \neq 2\lambda a^2$ the eigenproblem

$$A r^{(j)} = \mu_j r^{(j)}$$

has the eigenvalues

$$\mu_j = \delta t\, c + 2\lambda a^2 - \sqrt{4\lambda^2 a^4 - \gamma^2 b^2} \cos \frac{j\pi}{M}, \quad j = 1, \dots, M-1,$$

and the eigenvectors

$$r^{(j)} = \left(\left(\frac{2\lambda a^2 + \gamma b}{2\lambda a^2 - \gamma b} \right)^{\frac{1}{2}} \sin \frac{j\pi}{M}, \dots, \left(\frac{2\lambda a^2 + \gamma b}{2\lambda a^2 - \gamma b} \right)^{\frac{1}{2}} \sin \frac{(M-1) j\pi}{M} \right).$$

We now restrict problem (6.2) to a constant-coefficient problem. Then in the l^2 setting, we have the following stability result for the θ-schemes.

THEOREM. 6.11 *Let A be the matrix defined in Lemma 6.10. Assume that the eigenvalues of the symmetric part of this matrix, obtained by putting $b = 0$, are positive which corresponds to the inequality*

$$c > -\frac{4a^2}{(\delta x)^2}\sin^2\frac{\pi}{2M} \approx -a^2\pi^2, \tag{6.7}$$

where we have substituted $\delta x\, M = 1$, the size of the spatial domain $U = (0, 1)$.

If A is symmetric ($b = 0$) then the θ-scheme

$$(I + \theta A)W^{n+1} = (I - (1 - \theta)A)W^n$$

is stable in l^2-norm for

$$\theta \geq \frac{1}{2} - \frac{1}{4\lambda a^2 + \delta t\, c}.$$

For the nonsymmetric matrix A (with $b \neq 0$) the θ-scheme is stable if (6.7) *holds and $\theta \geq 1/2$.*

Proof. Taking into account equation (6.5) defining θ-schemes a sufficient condition for stability is

$$\big\|(I + \theta A)^{-1}\big(I - (1 - \theta)A\big)\big\| \leq 1.$$

Since

$$\big(I - (1 - \theta)A\big) = I + \theta A - \frac{1}{\theta}(I + \theta A) + \frac{1}{\theta}I = \left(1 - \frac{1}{\theta}\right)(I + \theta A) + \frac{1}{\theta}I,$$

we have

$$(I + \theta A)^{-1}\big(I - (1 - \theta)A\big) = \left(1 - \frac{1}{\theta}\right)I + \frac{1}{\theta}(I + \theta A)^{-1}.$$

Assume A is symmetric. Then also $(I + \theta A)^{-1}$ is symmetric, and the norm of

$$\left(1 - \frac{1}{\theta}\right)I + \frac{1}{\theta}(I + \theta A)^{-1}$$

is equal to its spectral radius. Then the estimate we are looking for can be reduced to

$$\max_{\mu_j}\left|\left(1 - \frac{1}{\theta}\right) + \frac{1}{\theta(1 + \theta\mu_j)}\right| \leq 1,$$

where μ_j are the eigenvalues of A.

By Lemma 6.10 we have

$$\mu_j = \delta t\, c + 2\lambda a^2 - 2\lambda a^2 \cos\frac{j\pi}{M} = \delta t\, c + 4\lambda a^2 \sin^2\frac{j\pi}{2M}, \quad j = 1, \ldots, M-1.$$

The condition

$$\left(1 - \frac{1}{\theta}\right) + \frac{1}{\theta(1+\theta\mu_j)} \le 1$$

gives

$$\theta - 1 + \frac{1}{1+\theta\mu_j} \le \theta \Rightarrow \theta\mu_j \ge 0 \Rightarrow \mu_j \ge 0,$$

which is (6.7).

The condition

$$\left(1 - \frac{1}{\theta}\right) + \frac{1}{\theta(1+\theta\mu_j)} \ge -1$$

gives

$$\begin{aligned}\theta - 1 + \frac{1}{1+\theta\mu_j} \ge -\theta \Rightarrow -1 + \frac{1}{1+\theta\mu_j} &\ge -2\theta \\ \Rightarrow \frac{\mu_j}{1+\theta\mu_j} \le 2 \Rightarrow \theta &\ge \frac{1}{2} - \frac{1}{\mu_j}.\end{aligned}$$

Taking into account the values of μ_j given by Lemma 6.10, we obtain

$$\max_j\left(-\frac{1}{\mu_j}\right) = -\frac{1}{\max_j \mu_j} \le -\frac{1}{\delta t\, c + 4\lambda a^2}.$$

Hence, if (6.7) holds and

$$\theta \ge \frac{1}{2} - \frac{1}{\delta t\, c + 4\lambda a^2}$$

the spectral radius is not greater than 1.

The stability for the nonsymmetric matrix A (with $b \neq 0$) follows from the estimate

$$\left\|(I+\theta A)^{-1}\big(I - (1-\theta)A\big)\right\| = \left\|\Big(1 - \frac{1}{\theta}\Big)I + \frac{1}{\theta}(I+\theta A)^{-1}\right\| \le 1.$$

Taking $W \in \mathbb{R}^{M+1}$ we get ($|\cdot|_2$ and $(\cdot,\cdot)$ denote the l^2 norm and the corresponding

scalar product in $\mathbb{R}^{M+1}$)

$$
\begin{aligned}
&\left|\left(\left(1-\frac{1}{\theta}\right)I+\frac{1}{\theta}(I+\theta A)^{-1}\right)W\right|_2^2\\
&\quad=\left(1-\frac{1}{\theta}\right)^2|W|_2^2+\frac{1}{\theta}\left(1-\frac{1}{\theta}\right)\left(W,(I+\theta A)^{-1}W\right)\\
&\qquad+\frac{1}{\theta}\left(1-\frac{1}{\theta}\right)\left((I+\theta A)^{-1}W,W\right)+\frac{1}{\theta^2}\left|(I+\theta A)^{-1}W\right|_2^2\\
&\quad=\left(\frac{1}{\theta}-1\right)^2|W|_2^2-\frac{1}{\theta}\left(\frac{1}{\theta}-1\right)\left((I+\theta A)(I+\theta A)^{-1}W,(I+\theta A)^{-1}W\right)\\
&\qquad-\frac{1}{\theta}\left(\frac{1}{\theta}-1\right)\left((I+\theta A)^{-1}W,(I+\theta A)(I+\theta A)^{-1}W\right)\\
&\qquad+\frac{1}{\theta^2}\left|(I+\theta A)^{-1}W\right|_2^2\\
&\quad=\left(\frac{1}{\theta}-1\right)^2|W|_2^2-\frac{1}{\theta}\left(\frac{1}{\theta}-1\right)\left((I+\theta A)^{-1}W,(I+\theta A)^{\top}(I+\theta A)^{-1}W\right)\\
&\qquad-\frac{1}{\theta}\left(\frac{1}{\theta}-1\right)\left((I+\theta A)^{-1}W,(I+\theta A)(I+\theta A)^{-1}W\right)\\
&\qquad+\frac{1}{\theta^2}\left|(I+\theta A)^{-1}W\right|_2^2\\
&\quad=\left(\frac{1}{\theta}-1\right)^2|W|_2^2+\left(\frac{2}{\theta}-\frac{1}{\theta^2}\right)\left|(I+\theta A)^{-1}W\right|_2^2\\
&\qquad-\left(\frac{1}{\theta}-1\right)\left((I+\theta A)^{-1}W,(A+A^{\top})(I+\theta A)^{-1}W\right).
\end{aligned}
$$

To complete the proof, we will show that the last expression is bounded by $|W|_2^2$.

This proof is in several steps. First, we prove that if all eigenvalues of the symmetric part of A are positive, then $(W,AW)\geq C|W|_2^2$ with $C>0$. Let A_0 denote the matrix A for $a^2=1$, $b=0$, $c=0$. By simple computations we get

$$(W,A_0W)=\lambda\sum_{i=0}^{M-1}(w_i-w_{i+1})^2,$$

where we have used the zero boundary conditions $w_0=w_M=0$.

On the other hand, expanding W in a series of eigenvectors of A_0 and using the positivity of eigenvalues of A_0, we get

$$(W,A_0W)\geq\min_j\mu_j|W|_2^2=4\lambda\sin^2\frac{\pi}{2M}|W|_2^2.$$

For the matrix A, we obtain similarly

$$(W, AW) = \lambda a^2 \sum_{i=0}^{M-1} (w_i - w_{i+1})^2 + c\delta t |W|_2^2$$
$$\geq \Big(c\delta t + 4\lambda a^2 \sin^2 \frac{\pi}{2M} \Big) |W|_2^2 \geq C|W|_2^2.$$

This proves that $(W, (I + \theta A)W) \geq C|W|_2^2$ with $C \geq 1$, which implies that $(I + \theta A)$ is invertible and gives

$$\big|(I + \theta A)^{-1} W\big|_2^2 \leq |W|_2^2.$$

The estimate for (W, AW) also shows that $(A + A^\top)$ is positive definite and

$$\Big(\frac{1}{\theta} - 1\Big)\big((I + \theta A)^{-1} W, (A + A^\top)(I + \theta A)^{-1} W\big) \geq 0.$$

Then we get

$$\left| \left(\Big(1 - \frac{1}{\theta}\Big) I + \frac{1}{\theta} (I + \theta A)^{-1} \right) W \right|_2^2 \leq \Big(\frac{1}{\theta} - 1\Big)^2 |W|_2^2 + \Big(\frac{2}{\theta} - \frac{1}{\theta^2}\Big) |W|_2^2 = |W|_2^2$$

since for $\theta \geq 1/2$ both coefficients on the right hand side are nonnegative and sum to 1. ∎

Remark. 6.1 *Let us observe that the above theorem gives for the heat equation the stability condition*

$$\theta \geq \frac{1}{2} - \frac{1}{4\lambda}$$

which is much stronger than the l^∞ stability condition of Theorem 6.8

$$\theta \geq 1 - \frac{1}{2\lambda}.$$

We will now carry the stability analysis of problem (6.2) with constant coefficients applying the discrete Fourier transform to the grid functions, which are solutions of (6.3). To simplify the presentation, we restrict considerations to stability with respect to initial data. Then we consider only an initial-value problem, and the domain U of our difference scheme is the whole $\mathbb{R}^d$.

Let $\delta x_1 = \delta x_2 = \cdots = \delta x_d = 1/M$. The grid points of $\mathbb{R}^d$ are indexed by vectors $\boldsymbol{k} \in \mathbb{Z}^d$ with components k_i, $i = 1, \ldots, d$. For a function $\{W_{\boldsymbol{k}}\}_{\boldsymbol{k} \in \mathbb{Z}^d}$ defined on the grid in $\mathbb{R}^d$ we introduce the discrete Fourier transform

$$\hat{W}(\xi) = \frac{1}{(2\pi)^{d/2}} \sum_{\boldsymbol{k} \in \mathbb{Z}^d} W_{\boldsymbol{k}} \mathrm{e}^{-i(\boldsymbol{k} \cdot \Delta x, \xi)} |\Delta x|,$$

where Δx is the vector with components $(\delta x_1, \delta x_2, \ldots, \delta x_d)$, $\boldsymbol{k} \cdot \Delta x$ denotes the vector with components $k_i \delta x_i$ and $|\Delta x| = \prod_{i=1}^{d} \delta x_i$.

The functions $\hat{W}$ are defined on the whole $\mathbb{R}^d$ and are periodic with period $2\pi M$ in the direction of each coordinate. Hence, we can restrict these functions to the cube $[-\pi M, \pi M]^d$ and get the inverse Fourier transform

$$W_{\boldsymbol{k}} = \frac{1}{(2\pi)^{d/2}} \int_{[-\pi M, \pi M]^d} \mathrm{e}^{i(\boldsymbol{k} \cdot \Delta x, \xi)} \hat{W}(\xi) d\xi.$$

We consider the grid functions W as elements of the space l^2 with the norm

$$|W|_2^2 = \sum_{\boldsymbol{k} \in \mathbb{Z}^d} \left| W_{\boldsymbol{k}} \right|^2 |\Delta x|,$$

Their Fourier transforms $\hat{W}$ belong to the space $L^2([-\pi M, \pi M]^d)$ with the norm

$$\|\hat{W}\|_2^2 = \frac{1}{(2\pi)^d} \int_{[-\pi M, \pi M]^d} \left| \hat{W}(\xi) \right|^2 d\xi.$$

Functions $\hat{W}(\xi)$ for a fixed ξ are from the Euclidean space $\mathbb{R}^d$ with the Euclidean norm denoted $|\cdot|$.

For the discrete Fourier transform holds also the Parseval identity known for the continuous Fourier transform

$$\|\hat{W}\|_2 = |W|_2.$$

Taking the discrete Fourier transform of the difference scheme

$$LW^{n+1} = RW^n \tag{6.8}$$

corresponding to the initial-value problem related to (6.2) with constant coefficients and zero free term, we obtain

$$\sum_{\boldsymbol{k}} \left(LW^{n+1}\right)_{\boldsymbol{k}} \mathrm{e}^{-i(\boldsymbol{k} \cdot \Delta x, \xi)} = \sum_{\boldsymbol{k}} \left(RW^n\right)_{\boldsymbol{k}} \mathrm{e}^{-i(\boldsymbol{k} \cdot \Delta x, \xi)}.$$

Since

$$\left(LW^{n+1}\right)_{\boldsymbol{k}} = \sum_{\boldsymbol{j}} l_{\boldsymbol{j}} W^{n+1}_{\boldsymbol{k}-\boldsymbol{j}}, \quad \left(RW^n\right)_{\boldsymbol{k}} = \sum_{\boldsymbol{j}} r_{\boldsymbol{j}} W^n_{\boldsymbol{k}-\boldsymbol{j}}$$

equation (6.8) leads to the equation for Fourier's transforms

$$\hat{L}(\xi) \hat{W}^{n+1}(\xi) = \hat{R}(\xi) \hat{W}^n(\xi),$$

where

$$\hat{L}(\xi) = \sum_{\boldsymbol{j}} l_{\boldsymbol{j}} \mathrm{e}^{-i(\boldsymbol{j}\cdot\Delta x,\xi)}, \quad \hat{R}(\xi) = \sum_{\boldsymbol{j}} r_{\boldsymbol{j}} \mathrm{e}^{-i(\boldsymbol{j}\cdot\Delta x,\xi)}.$$

The matrix

$$G(\xi) = \hat{L}^{-1}(\xi)\hat{R}(\xi)$$

is called the *amplification matrix* and is a continuous function of ξ. Then the equation for Fourier's transforms reads

$$\hat{W}^n(\xi) = \big(G(\xi)\big)^n \hat{W}^0(\xi).$$

THEOREM. 6.12 (von Neumann stability condition) *A necessary condition for stability in the norm l^2 with respect to initial data for the problem* (6.2) *with constant coefficients is the existence of a constant K such that the spectral radius of $G(\xi)$ has the estimate*

$$\rho\big(G(\xi)\big) \leq 1 + K\delta t, \quad \forall \xi.$$

Since the spectral radius of A is the greatest absolute value of the eigenvalues of A, we can reformulate the above estimate in terms of the eigenvalues of $G(\xi)$

$$\forall \xi \quad |\mu_j(\xi)| \leq 1 + K\delta t,$$

where $\mu_j(\xi)$ denotes any eigenvalue of the amplification matrix $G(\xi)$.

Proof. Due to Lemma 6.9 we have the estimate

$$\rho(A^n) \leq \|A^n\| \leq \|A\|^n,$$

where $\|A\|$ denotes the operator norm of A.

Let us now assume that the necessary condition is not satisfied. Then for each K there exist a vector ξ_K and an eigenvalue $\mu(\xi_K)$ of $G(\xi_K)$ such that $|\mu(\xi_K)| > 1 + K\delta t$. Thus for $T = n\delta t$ we have

$$|\mu(\xi_K)|^n > 1 + Kn\delta t = 1 + KT \;\; \Rightarrow \;\; \|G^n(\xi_K)\| > 1 + KT,$$

where $\|G^n(\xi_K)\|$ is the operator norm of matrix $G^n(\xi_K)$ in $\mathbb{R}^d$.

Since $G^n(\xi)$ is a continuous function, there exists an open neighborhood I_K of ξ_K with a positive volume, where we have the estimate $\|G^n(\xi)\| > 1 + KT$, for $\xi \in I_K$. We can now select an initial condition W_K^0 such that $\hat{W}_K^0$ is zero in the exterior of I_K and $|G^n(\xi)\hat{W}_K^0(\xi)| > (1 + KT)|\hat{W}_K^0(\xi)|$ for $\xi \in I_K$. Since the volume of I_K is positive we get the estimate with smaller but still unbounded from above constant K_1

$$\|G^n\hat{W}_K^0\|_2 > (1 + K_1T)\|\hat{W}_K^0\|_2.$$

By the Parseval identity and the above inequality, we have

$$\sup_{W^0} \frac{|W^n|_2}{|W^0|_2} = \sup_{\hat{W}^0} \frac{\|\hat{W}^n\|_2}{\|\hat{W}^0\|_2} \geq \frac{\|G^n \hat{W}_K^0\|_2}{\|\hat{W}_K^0\|_2} > (1 + K_1 T),$$

which contradicts stability, as K_1 is arbitrarily large. ■

COROLLARY. 6.13 *The proof of the above theorem shows that the sufficient condition for stability is*

$$\sup_{\xi} \|G^n(\xi)\| \leq C,$$

for each n such that $n\delta t \leq T$ and $C > 0$ independent of n.

The von Neumann stability condition for an initial value problem is also a necessary condition for stability of an initial-boundary value problem. The extension of the von Neumann stability analysis to a sufficient condition for initial-boundary value problems is complicated and is beyond the scope of these lecture notes.

But initial-boundary value problems in a rectangular domain, say $U = (-1, 1)^d$, with periodic boundary data, can be easily reduced to the already discussed initial value problems. Let as previously $\delta x_1 = \delta x_2 = \cdots = \delta x_d = 1/M$. The grid point of J_U are indexed by the vectors

$$\boldsymbol{k} \in Z_M = \{\boldsymbol{k} = (k_1, \ldots, k_d)\colon\ k_i = 0, \pm 1, \pm 2, \ldots, \pm M\}.$$

A grid function $\{W_{\boldsymbol{k}}\}_{\boldsymbol{k} \in Z_M}$ defined on the discrete grid J_U can be expanded in a series of trigonometric polynomials

$$W_{\boldsymbol{k}} = W(\boldsymbol{k} \cdot \Delta x) = \frac{1}{(2\pi)^{d/2}} \sum_{\boldsymbol{l} \in Z_M}{}' \hat{W}(\xi_{\boldsymbol{l}}) \mathrm{e}^{-i(\boldsymbol{k} \cdot \Delta x, \xi_{\boldsymbol{l}})}, \text{ where } \xi_{\boldsymbol{l}} = \boldsymbol{l}\pi,$$

where the prime in the summation sign indicates that the summation has to be modified for $l_i = \pm M$ to take into account the periodicity of boundary conditions.

The coefficients $\hat{W}(\xi_{\boldsymbol{l}})$ are defined by the *finite Fourier transform*

$$\hat{W}(\xi_{\boldsymbol{l}}) = \frac{1}{(2\pi)^{d/2}} \sum_{\boldsymbol{k} \in Z_M}{}' W_{\boldsymbol{k}} \mathrm{e}^{-i(\boldsymbol{k} \cdot \Delta x, \xi_{\boldsymbol{l}})} |\Delta x|, \text{ for } \xi_{\boldsymbol{l}} = \boldsymbol{l}\pi.$$

Functions $\hat{W}$ belong to the space l^2 with the norm

$$|\hat{W}|_2^2 = \sum_{\boldsymbol{l} \in Z_M}{}' |\hat{W}(\xi_{\boldsymbol{l}})|^2,$$

whereas functions W belong to the space l^2 with the norm

$$|W|_2^2 = \sum_{\boldsymbol{k}\in Z_M}{}' |W_{\boldsymbol{k}}|^2 |\Delta x|.$$

With the above norms we have the Parseval identity $|W|_2 = |\hat{W}|_2$. Similarly like for initial value problems we can define the amplification matrix $G = G(\xi_{\boldsymbol{l}})$ which is now defined for a discrete set of points $\xi_{\boldsymbol{l}}$. For this amplification matrix the von Neumann stability condition (Theorem 6.12 and Corollary 6.13) remains valid but the proof of Theorem 6.12 requires some modifications.

Proof of Theorem 6.12 (for the finite Fourier transform). Since G^n is an operator of multiplication by a matrix, we can compute its norm by applying a well-known property of operators of multiplication. If B is an operator of multiplication by a function $b(x)$ in spaces L^2 (or l^2), then the operator norm of B is given by the formula

$$\|B\| = \sup_{u\in L^2} \frac{\left(\int |b(x)u(x)|^2 dx\right)^{\frac{1}{2}}}{\left(\int |u(x)|^2 dx\right)^{\frac{1}{2}}} = \sup_x |b(x)|.$$

By the Parseval identity and the above property of multiplication operators, we have

$$\sup_{W^0} \frac{|W^n|_2}{|W^0|_2} = \sup_{\hat{W}^0} \frac{|\hat{W}^n|_2}{|\hat{W}^0|_2} = \sup_{\hat{W}^0} \frac{|G^n \hat{W}^0|_2}{|\hat{W}^0|_2} = \sup_{\boldsymbol{l}} \|\big(G(\xi_{\boldsymbol{l}})\big)^n\|.$$

Since the stability requirement is the boundedness of $G^n(\xi_{\boldsymbol{l}})$ for all $\boldsymbol{l} \in Z_M$ and

$$\|\big(G(\xi_{\boldsymbol{l}})\big)^n\| \geq \rho\big(G^n(\xi_{\boldsymbol{l}})\big) = \big(\rho\big(G(\xi_{\boldsymbol{l}})\big)\big)^n,$$

then a necessary condition for stability is the existence of a constant $K > 1$ such that

$$\big(\rho\big(G(\xi_{\boldsymbol{l}})\big)\big)^n < K \quad \forall \boldsymbol{l} \in Z_M.$$

Taking into account that $n\delta t \leq T$ we obtain

$$\rho\big(G(\xi_{\boldsymbol{l}})\big) \leq K^{\delta t/T} \leq 1 + K_1 \delta t \quad \forall \boldsymbol{l} \in Z_M,$$

which completes the proof. ■

It can be shown that if G is a normal matrix, then its operator norm is equal to its spectral radius. Then the von Neumann condition is also sufficient, in particular, it is sufficient for one-dimensional problems. It appears, however, that even when the condition is sufficient the constant in the condition can lead to the rapid growth of a grid solution. The following example illustrates the problem.

Example. 6.14 *Consider the convection-diffusion equation*

$$\frac{\partial u}{\partial t} + b\frac{\partial u}{\partial x} = a^2\frac{\partial^2 u}{\partial x^2}.$$

Applying to the Euler explicit scheme

$$\frac{w_k^{n+1} - w_k^n}{\delta t} + b\frac{w_{k+1}^n - w_{k-1}^n}{2\delta x} = a^2\frac{w_{k+1}^n - 2w_k^n + w_{k-1}^n}{(\delta x)^2}$$

the discrete Fourier transform, we obtain the amplification function ($d = 1$, hence G is one-dimensional)

$$G(\xi) = 1 - 2\kappa + (\kappa + \nu/2)e^{-i\delta x\,\xi} + (\kappa - \nu/2)e^{i\delta x\,\xi},$$

where $\kappa = a^2\frac{\delta t}{(\delta x)^2}$, $\nu = b\frac{\delta t}{\delta x}$. Denoting $s = \sin(\delta x\,\xi/2)$ we get

$$\begin{aligned} G(\xi) &= 1 - 2\kappa + 2\kappa\cos(\delta x\,\xi) - i\nu\sin(\delta x\,\xi), \\ |G(\xi)|^2 &= (1 - 4\kappa s^2)^2 + 4\nu^2 s^2(1 - s^2). \end{aligned}$$

Taking $s^2 = 1$ we obtain $|G| \le 1$ for $\kappa \le \frac{1}{2}$, which is the necessary condition for stability. As $\nu^2 = b^2\left(\frac{\delta t}{\delta x}\right)^2 = (b^2/a^2)\kappa\delta t$ and $\max_{s^2\in[0,1]} s^2(1 - s^2) = \frac{1}{4}$ we get for $\kappa \le \frac{1}{2}$ the estimate

$$|G(\xi)|^2 = (1-4\kappa s^2)^2+4\nu^2 s^2(1-s^2) \le 1+4\nu^2 s^2(1-s^2) \le 1+\nu^2 \le 1+\frac{b^2}{2a^2}\delta t,$$

so that the von Neumann condition is satisfied and, since the problem is one-dimensional, the condition is sufficient for stability. On the other hand, taking $\nu = 1$, $\kappa = \frac{1}{4}$ and $s^2 = \frac{1}{2}$, we obtain $|G|^2 = \frac{5}{4}$, which gives a rapid growth of Fourier's modes.

The above example suggests the introduction of a stronger notion of stability.

DEFINITION. 6.15 *A two-time-level difference scheme is said to be* strongly stable *in the norm l^2 with respect to initial conditions if the following estimate holds for the amplification matrix*

$$\forall\xi \quad \|G(\xi)\| \le 1.$$

In Example 6.14 the condition $|G(\xi)| \le 1$ gives

$$\nu^2 \le 2\kappa \le 1. \tag{6.9}$$

Indeed, from the expression

$$|G(\xi)|^2 = (1 - 4\kappa s^2)^2 + 4\nu^2 s^2(1 - s^2),$$

we obtain for $s^2 = 1$ the estimate $2\kappa \leq 1$. Rewriting $|G(\xi)|^2$ as

$$|G(\xi)|^2 = 1 - 4s^2(2\kappa - \nu^2) + 4s^4(4\kappa^2 - \nu^2)$$

we obtain $|G(\xi)| \leq 1$ provided that

$$4s^4(4\kappa^2 - \nu^2) \leq 4s^2(2\kappa - \nu^2).$$

Thus for $s^2 > 0$ we get

$$s^2(4\kappa^2 - \nu^2) \leq (2\kappa - \nu^2).$$

Excluding the case $2\kappa = \nu^2 = 1$ and taking the limit $s^2 \to 0$ we obtain from the above inequality $2\kappa - \nu^2 \geq 0$. Since the excluded case falls into the general inequality, it proves (6.9).

Number $\frac{\nu}{\kappa}$ is called the *Peclet mesh number* and relates the convection term $b\frac{\partial u}{\partial x}$ to the diffusion term $a^2\frac{\partial^2 u}{\partial x^2}$. The condition $\frac{\nu^2}{\kappa} \leq 2$ means that the convection term cannot be too large with respect to the diffusion term. The maximal admissible $\kappa = \frac{1}{2}$ restricts the Peclet mesh number $\frac{\nu}{\kappa} \leq 2$ which implies $\delta x \leq \frac{2a^2}{b}$. The violation of this restriction of the grid step can generate spurious oscillations of the scheme.

We can improve the stability by taking an *upwind scheme*, which applies a different approximation of the first-order x derivative. Let us consider the equation of Example 6.14

$$\frac{\partial u}{\partial t} + b\frac{\partial u}{\partial x} = a^2\frac{\partial^2 u}{\partial x^2}.$$

If $b > 0$ the left hand side of the equation is a first-order operator with the traveling wave solution $u(t, x) = F(x - bt)$, where $F(\cdot) = g(\cdot)$ is given by the initial condition. We can say that the profile $F(\cdot)$ drifts in the positive x direction ("wind blows to the right"). Looking from the node (k, n) the node $(k-1, n)$ is "upwind". Applying that approximation of the first-order derivative we obtain for the Euler explicit scheme

$$\frac{w_k^{n+1} - w_k^n}{\delta t} + b\frac{w_k^n - w_{k-1}^n}{\delta x} = a^2\frac{w_{k+1}^n - 2w_k^n + w_{k-1}^n}{(\delta x)^2}$$

the amplification function

$$G(\xi) = 1 - 2\kappa - \nu + (\kappa + \nu)e^{-i\delta x\,\xi} + \kappa e^{i\delta x\,\xi}.$$

After simplifications we get

$$G(\xi) = 1 - 2\kappa - \nu + 2\kappa\cos(\delta x\,\xi) + \nu\cos(\delta x\,\xi) - i\nu\sin(\delta x\,\xi),$$
$$|G(\xi)|^2 = (1 - 2(2\kappa+\nu)s^2)^2 + 4\nu^2 s^2(1-s^2).$$

The condition of strong stability $|G(\xi)| \le 1$ gives the inequality

$$2\kappa + \nu \le 1. \tag{6.10}$$

To prove that inequality, observe that condition $|G(\xi)|^2 \le 1$ gives $2\kappa+\nu \le 1$ for $s^2 = 1$. Since $0 < \nu^2 \le \nu \le 2\kappa+\nu$ for $b > 0$, then for $2\kappa+\nu \le 1$ and arbitrary $s^2 \in [0,1]$ we have

$$\begin{aligned}|G(\xi)|^2 &= (1 - 2(2\kappa+\nu)s^2)^2 + 4\nu^2 s^2(1-s^2)\\ &\le (1 - 2(2\kappa+\nu)s^2)^2 + 4(2\kappa+\nu)s^2(1-s^2)\\ &= 1 + 4s^4(2\kappa+\nu)(2\kappa+\nu-1) \le 1,\end{aligned}$$

so the condition of strong stability holds.

Inequality (6.10) is called the **CFL** condition (Courant, Friedrichs and Lewy [12]) and for small κ is less restrictive than (6.9). For example, taking $\kappa = \frac{1}{50}$ we obtain from inequality (6.10) $\nu \le 0.96$ and $\frac{\nu^2}{\kappa} \approx 46$, a value much larger than $\frac{\nu^2}{\kappa} \le 2$ permitted by inequality (6.9).

Remark. 6.2 *The von Neumann stability condition applies only to constant-coefficient equations where the Fourier transform is well defined. Nevertheless, the von Neumann analysis is also used for variable-coefficient equations "freezing" coefficients in their constant values. Since stability is a local property "freezing" works well for many variable-coefficient equations.*

6.5 The Black-Scholes equation in the original variables

We return now to the Black-Scholes equation in a general setting of variable coefficients

$$\frac{\partial V(t,s)}{\partial t} - \frac{1}{2}\sigma^2(t,s)s^2\frac{\partial^2 V(t,s)}{\partial s^2} - r(t)s\frac{\partial V(t,s)}{\partial s} + r(t)V(t,s) = 0, \tag{6.11}$$

with the initial condition

$$V(0,s) = V_0(s).$$

Ignoring the unboundedness of the coefficients we write the θ-scheme for that equation

$$(I + \theta A^{n+1})W^{n+1} = \big(I - (1-\theta)A^n\big)W^n, \tag{6.12}$$

where

$$A^n_{k,k} = (k\sigma^n_k)^2 + r^n,\ A^n_{k,k+1} = -\frac{1}{2}\big((k\sigma^n_k)^2 + kr^n\big),\ A_{k,k-1} = -\frac{1}{2}\big((k\sigma^n_k)^2 - kr^n\big).$$

Numerical experiments show that the scheme is unconditionally convergent for $\theta \geq 1/2$, i.e., for the Euler implicit and the Crank-Nicolson schemes, independently from the fact that the coefficients are unbounded

Explanation

For (6.11) we can carry on a similar analysis as for problem (5.11). In particular, using similar methods, we can prove the existence of weak solutions to (6.11). We have only to replace the space $H^1_0(U)$ by the space

$$\mathcal{V} = \{u \in L^2(\mathbb{R}_+):\ x\frac{du}{dx} \in L^2(\mathbb{R}_+)\}$$

with the norm

$$\|u\|_{\mathcal{V}} = \Big\|x\frac{du}{dx}\Big\|_{L^2(\mathbb{R}_+)},$$

where the derivative $\frac{du}{dx}$ is understood in a weak sense.

DEFINITION. 6.16 *$V(t,x)$ such that $V \in C([0,T]; L^2(\mathbb{R}_+)) \cap L^2(0,T;\mathcal{V})$ and $\frac{dV}{dt} \in L^2(0,T;\mathcal{V}')$, where $\mathcal{V}'$ is the dual space to $\mathcal{V}$, is called the weak solution of the Black-Scholes equation* (6.11)*, if for each $u \in \mathcal{V}$*

$$\begin{aligned} &\Big(\tfrac{\partial V}{\partial t}(t,\cdot), u\Big) + B^t[V(t,\cdot), u] = 0, \quad t \in (0,T], \\ &V(0,\cdot) = V_0, \ \ \textit{on}\ \ \mathbb{R}_+, \end{aligned} \tag{6.13}$$

where

$$\begin{aligned} B^t[v,u] = &\int_{\mathbb{R}_+} \frac{1}{2}s^2\sigma^2(t,s)\frac{\partial v}{\partial s}\frac{\partial u}{\partial s}ds \\ &+ \int_{\mathbb{R}_+}\left(-r(t) + \sigma^2(t,s) + s\sigma(t,s)\frac{\partial \sigma}{\partial s}\right)s\frac{\partial v}{\partial s}uds + r(t)\int_{\mathbb{R}_+} vuds. \end{aligned}$$

An energy estimate is crucial for the existence of weak solutions. In the case of the Black-Scholes equation (6.11), we have the following energy estimate.

THEOREM. 6.17 (Energy estimate for (6.11)**)** *Assume*

1. $0 < \sigma_L \leq \sigma(t,s) \leq \sigma_U$, *where* σ_L *and* σ_U *are constants.*

2. *There is a constant* C_σ *such that*

$$\left| s\frac{\partial \sigma}{\partial s} \right| \leq C_\sigma, \quad \forall t \in [0,T],\ \forall s \in \mathbb{R}_+.$$

Then

$$\begin{aligned} &\exists \alpha > 0, \quad |B^t[v,u]| \leq \alpha \|v\|_{\mathcal{V}} \|u\|_{\mathcal{V}}, \\ &\exists \gamma > 0, \quad \frac{\sigma_L^2}{4}\|v\|_{\mathcal{V}}^2 \leq B^t[v,v] + \gamma \|v\|_{L^2(\mathbb{R}_+)}^2. \end{aligned}$$

The proof of this energy estimate is similar to the proof of Theorem 5.20 if we use the $\mathbb{R}^d$ version of Poincaré's inequality

$$\forall v \in \mathcal{V}, \quad \|v\|_{L^2(\mathbb{R}_+)} \leq 2\left\| x\frac{dv}{dx} \right\|_{L^2(\mathbb{R}_+)},$$

which can be obtained from the integral identity

$$2\int_{\mathbb{R}_+} xv(x)\frac{dv}{dx}dx = -\int_{\mathbb{R}_+} v^2(x)dx.$$

By the above energy estimate, we obtain the following existence theorem.

THEOREM. 6.18 *If* $V_0 \in L^2(\mathbb{R}_+)$ *and assumptions 1. and 2. of Theorem 6.17 are fulfilled, then there exists a unique weak solution of problem* (6.13) *and for each* t, $0 < t \leq T$, *we have the estimate*

$$e^{-2\gamma t}\|V(t,\cdot)\|_{L^2(\mathbb{R}_+)}^2 + \frac{1}{2}\sigma_L^2 \int_0^t e^{-2\gamma\tau}\|V(\tau,\cdot)\|_{\mathcal{V}}^2 d\tau \leq \|V_0\|_{L^2(\mathbb{R}_+)}^2,$$

where γ *is the constant from Theorem 6.17.*

The estimate of Theorem 6.18 explains why despite unbounded coefficients the scheme (6.12) possesses good numerical properties.

6.6 Finite differences in many dimensions

Generalizations of the one-dimensional methods

We are going to extend the one-dimensional finite difference schemes, we have been studying, to many dimensions. Before we proceed, we would like to emphasize that the approach we will use is the same as in the case of the one-dimensional schemes. The concepts of consistency, stability, and convergence remain valid as they have been formulated for multi-dimensional finite differences. Our basic tools will be the Lax equivalence theorem (Theorem 6.5) and the von Neumann analysis of stability.

We begin with the Dirichlet problem for the two-dimensional heat equation in the rectangular domain $U = [0,1]^2$

$$\begin{aligned} &\frac{\partial u}{\partial t} - a^2\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = 0, \quad (t,x,y) \in (0,T] \times U, \\ &u(t,x,y) = q(t,x,y), \quad (x,y) \in \partial U,\ t \in [0,T], \\ &u(0,x,y) = g(x,y), \quad (x,y) \in U. \end{aligned} \tag{6.14}$$

In the analogy to the one-dimensional approach, we introduce in U a rectangular grid with spacing $\delta x \times \delta y$ where

$$\delta x = \frac{1}{M_x}, \quad \delta y = \frac{1}{M_y},$$

and divide the time interval into N subintervals of length $\delta t = \frac{T}{N}$. The grid points of this mesh are denoted

$$\begin{aligned} (x_j, y_k) =& (j \cdot \delta x, k \cdot \delta y),\ j = 0, \ldots, M_x,\ k = 0, \ldots, M_y, \\ & t_n = n \cdot \delta t,\ n = 0, \ldots, N. \end{aligned}$$

The time derivative in the heat equation is approximated by the forward difference

$$\frac{\partial u(t_n, x_j, y_k)}{\partial t} \approx \frac{u(t_{n+1}, x_j, y_k) - u(t_n, x_j, y_k)}{\delta t},$$

and the second-order space derivatives are approximated by the central differences

$$\begin{aligned} \frac{\partial^2 u(t_n, x_j, y_k)}{\partial x^2} &\approx \frac{u(t_n, x_{j+1}, y_k) - 2u(t_n, x_j, y_k) + u(t_n, x_{j-1}, y_k)}{(\delta x)^2}, \\ \frac{\partial^2 u(t_n, x_j, y_k)}{\partial y^2} &\approx \frac{u(t_n, x_j, y_{k+1}) - 2u(t_n, x_j, y_k) + u(t_n, x_j, y_{k-1})}{(\delta y)^2}. \end{aligned}$$

We denote by $w_{j,k}^n$ the approximation of $u(t_n, x_j, y_k)$. To simplify the notation, we define several finite difference operators. We begin with the first-order difference operator

$$\delta_x w_{j,k}^n = \frac{w_{j+\frac{1}{2},k}^n - w_{j-\frac{1}{2},k}^n}{\delta x}.$$

Since points $x_j \pm \frac{1}{2}\delta x$ are not grid points we use in fact the central differences

$$\delta_x w_{j,k}^n = \frac{w_{j+1,k}^n - w_{j-1,k}^n}{2\delta x}.$$

That leads to the following second-order difference operator

$$\delta_{xx} w_{j,k}^n = \frac{w_{j+1,k}^n - 2w_{j,k}^n + w_{j-1,k}^n}{(\delta x)^2}.$$

Analogously, we define the operators δ_y and δ_{yy}.

We begin with the two-dimensional explicit Euler scheme

$$\frac{w_{j,k}^{n+1} - w_{j,k}^n}{\delta t} - a^2(\delta_{xx} + \delta_{yy})w_{j,k}^n = 0.$$

Solving for $w_{j,k}^{n+1}$ we obtain

$$\begin{aligned} w_{j,k}^{n+1} =& a^2\lambda_x\big(w_{j-1,k}^n + w_{j+1,k}^n\big) + a^2\lambda_y\big(w_{j,k-1}^n + w_{j,k+1}^n\big) \\ &+ (1 - 2a^2\lambda_x - 2a^2\lambda_y)w_{j,k}^n, \end{aligned}$$

where

$$\lambda_x = \frac{\delta t}{(\delta x)^2}, \quad \lambda_y = \frac{\delta t}{(\delta y)^2}.$$

By the above equation we can compute $w_{j,k}^{n+1}$ at all internal points (x_j, y_k), $j = 1, \ldots, M_x - 1$, $k = 1, \ldots, M_y - 1$.

The boundary conditions

$$\begin{aligned} &w_{j,k}^{n+1} = g_0(t_{n+1}, x_j, y_k), \\ &\qquad \text{for } (j,k) \in \{0, M_x\} \times \{0, \ldots, M_y\} \cup \{0, \ldots, M_x\} \times \{0, M_y\} \end{aligned}$$

supplement the solution on ∂U.

The truncation error of the scheme is $O(\delta t) + O\big(|\delta x|^2 + |\delta y|^2\big)$ and can be computed similarly like in the proof of Theorem 6.7. Performing the von Neumann stability analysis we obtain the following amplification function

$$\begin{aligned} G(\xi_x, \xi_y) &= 1 + a^2\lambda_x\big(\mathrm{e}^{-i\xi_x} - 2 + \mathrm{e}^{i\xi_x}\big) + a^2\lambda_y\big(\mathrm{e}^{-i\xi_y} - 2 + \mathrm{e}^{i\xi_y}\big) \\ &= 1 - 4a^2\lambda_x \sin^2\frac{\xi_x}{2} - 4a^2\lambda_y \sin^2\frac{\xi_y}{2}. \end{aligned}$$

The maximum of G occurs at $(0,0)$ and is equal 1 and the minimum, at (π,π) and is equal $1-4a^2(\lambda_x+\lambda_y)$. Thus $|G(\xi_x,\xi_y)|\leq 1$ if

$$1-4a^2(\lambda_x+\lambda_y)\geq -1 \implies \lambda_x+\lambda_y\leq \frac{1}{2a^2}.$$

Hence, we have obtained a numerical method easy to implement and with a low computational complexity but with a restrictive stability condition. That is not a surprise as we have used an explicit scheme.

To improve the stability, we apply the Crank-Nicolson scheme

$$\frac{w_{j,k}^{n+1}-w_{j,k}^n}{\delta t}-\frac{a^2}{2}(\delta_{xx}+\delta_{yy})\big(w_{j,k}^{n+1}+w_{j,k}^n\big)=0.$$

Solving for $w_{j,k}^{n+1}$ we obtain

$$\Big(1-\frac{\delta t}{2}a^2(\delta_{xx}+\delta_{yy})\Big)w_{j,k}^{n+1}=\Big(1+\frac{\delta t}{2}a^2(\delta_{xx}+\delta_{yy})\Big)w_{j,k}^n.$$

Introducing the matrix

$$W^n=\big(w_{j,k}^n\big), j=1,\ldots,M_x-1, k=1,\ldots,M_y-1$$

we can write the above equation as

$$\Big(I-\frac{1}{2}C\Big)W^{n+1}=\Big(I+\frac{1}{2}C\Big)W^n$$

where C is a $(M_y-1)\times(M_y-1)$ tridiagonal block matrix

$$C=\begin{pmatrix} D_x & D_y & & \\ D_y & D_x & D_y & \\ & \ddots & \ddots & \\ & & D_y & D_x \end{pmatrix}$$

with the matrix entries of dimension $(M_x-1)\times(M_x-1)$

$$D_x=\begin{pmatrix} \alpha & \beta & & \\ \beta & \alpha & \beta & \\ & \ddots & \ddots & \\ & & \beta & \alpha \end{pmatrix},\quad D_y=\lambda_y a^2 I,$$

where

$$\alpha=-2a^2\big(\lambda_x+\lambda_y\big),\quad \beta=\lambda_x a^2.$$

The truncation error of the scheme is $O\big(|\delta t|^2 + |\delta x|^2 + |\delta y|^2\big)$ which can be computed like in Theorem 6.7. The von Neumann stability analysis gives

$$\begin{aligned}\big(1+2a^2\lambda_x \sin^2 \frac{\xi_x}{2} + 2a^2\lambda_y \sin^2 \frac{\xi_y}{2}\big)G(\xi_x,\xi_y)\\ = \big(1 - 2a^2\lambda_x \sin^2 \frac{\xi_x}{2} - 2a^2\lambda_y \sin^2 \frac{\xi_y}{2}\big).\end{aligned}$$

Since $\left|\frac{1-z}{1+z}\right| \leq 1$ for all $z \geq 0$ then $|G(\xi_x,\xi_y)| \leq 1$ and the scheme is unconditionally stable.

Thus we have obtained a scheme with good approximation and stability properties but very costly in implementation. In each time step we have to solve a linear system of $(M_x - 1) \times (M_y - 1)$ equations. Since the system is not tridiagonal its solution is very expensive. The fact that the system is block-tridiagonal can reduce the cost, but this reduction is minor.

Alternating direction method

A significant reduction of computational complexity can be achieved by splitting the two-dimensional problem into a sequence of one-dimensional problems. Before passing to rigorous mathematical considerations, let us describe informally the idea behind the time-splitting method.

Consider the semi-discrete equation

$$\frac{dW}{dt} = \Lambda W,$$

where Λ is a discrete matrix representation of a differential operator.

The solution to this equation can be written formally as

$$W(t) = \exp(\Lambda t)W(0).$$

Assume that Λ can be split into

$$\Lambda = \Lambda_1 + \Lambda_2.$$

We can think about splitting with Λ_1 corresponding to x derivatives and Λ_2, to y derivatives, but, in fact, the splitting can be quite general.

We want to apply the above splitting to replace $\mathrm{e}^{(\Lambda_1+\Lambda_2)t}$ by $\mathrm{e}^{\Lambda_1 t}\mathrm{e}^{\Lambda_2 t}$. These two operators are equal only if Λ_1 and Λ_2 commute ($\Lambda_1\Lambda_2 = \Lambda_2\Lambda_1$). However, when the time increment is small then also the error generated by substituting $\mathrm{e}^{(\Lambda_1+\Lambda_2)t}$ by $\mathrm{e}^{\Lambda_1 t}\mathrm{e}^{\Lambda_2 t}$ is small.

Applying the above splitting to a multi-dimensional finite difference scheme with the introduced earlier time discretization, we have by Taylor's expansion

$$W^{n+1} = \mathrm{e}^{\delta t(\Lambda_1+\Lambda_2)}W^n = \Big(\mathrm{e}^{\delta t\Lambda_1}\mathrm{e}^{\delta t\Lambda_2} + \frac{(\delta t)^2}{2}(\Lambda_2\Lambda_1 - \Lambda_1\Lambda_2) + O\big((\delta t)^3\big)\Big)W^n. \tag{6.15}$$

Hence, up to the local error term $O(|\delta t|^2)$ we can replace $\mathrm{e}^{\delta t(\Lambda_1+\Lambda_2)}$ by $\mathrm{e}^{\delta t\Lambda_1}\mathrm{e}^{\delta t\Lambda_2}$.

The idea of operator splitting is the basis of a very powerful method that is especially used for solving parabolic equations. This method is called the *alternating direction implicit* (ADI) method. We will present various ADI algorithms for the already introduced Dirichlet problem for the two-dimensional heat equation.

Peaceman-Rachford scheme. This scheme is obtained by splitting the Crank-Nicolson scheme

$$\Big(1 - \frac{\delta t}{2}a^2(\delta_{xx} + \delta_{yy})\Big)w_{j,k}^{n+1} = \Big(1 + \frac{\delta t}{2}a^2(\delta_{xx} + \delta_{yy})\Big)w_{j,k}^{n}. \tag{6.16}$$

Since

$$1 \pm \frac{\delta t}{2}a^2\delta_{xx} \pm \frac{\delta t}{2}a^2\delta_{yy} = \Big(1 \pm \frac{\delta t}{2}a^2\delta_{xx}\Big)\Big(1 \pm \frac{\delta t}{2}a^2\delta_{yy}\Big) - \frac{(\delta t)^2}{4}a^4\delta_{xx}\delta_{yy}$$

we can write (6.16) as

$$\begin{aligned}\Big(1 - \frac{\delta t}{2}a^2\delta_{xx}\Big)\Big(1 - \frac{\delta t}{2}a^2\delta_{yy}\Big)w_{j,k}^{n+1} &= \Big(1 + \frac{\delta t}{2}a^2\delta_{xx}\Big)\Big(1 + \frac{\delta t}{2}a^2\delta_{yy}\Big)w_{j,k}^{n} \\ &\quad + \frac{(\delta t)^2}{4}a^4\delta_{xx}\delta_{yy}\big(w_{j,k}^{n+1} - w_{j,k}^{n}\big).\end{aligned}$$

By the Taylor expansion we have $w_{j,k}^{n+1} = w_{j,k}^{n} + O(\delta t)$. Thus the term

$$\frac{(\delta t)^2}{4}a^4\delta_{xx}\delta_{yy}\big(w_{j,k}^{n+1} - w_{j,k}^{n}\big)$$

is of order $O\big(|\delta t|^3\big)$ that is the order of the discretization error.

Neglecting this term as we neglect the discretization error, we arrive at the scheme

$$\Big(1 - \frac{\delta t}{2}a^2\delta_{xx}\Big)\Big(1 - \frac{\delta t}{2}a^2\delta_{yy}\Big)w_{j,k}^{n+1} = \Big(1 + \frac{\delta t}{2}a^2\delta_{xx}\Big)\Big(1 + \frac{\delta t}{2}a^2\delta_{yy}\Big)w_{j,k}^{n}. \tag{6.17}$$

That scheme is up to $O\big(|\delta t|^3\big)$ equivalent to the Crank-Nicolson scheme. In this scheme, however, the finite differences in the directions of x and y are separated

and introducing the intermediate level $W^{n+\frac{1}{2}}$ we get

$$\Big(1-\frac{\delta t}{2}a^2\delta_{xx}\Big)w_{j,k}^{n+\frac{1}{2}}=\Big(1+\frac{\delta t}{2}a^2\delta_{yy}\Big)w_{j,k}^{n}, \tag{6.18}$$

$$\Big(1-\frac{\delta t}{2}a^2\delta_{yy}\Big)w_{j,k}^{n+1}=\Big(1+\frac{\delta t}{2}a^2\delta_{xx}\Big)w_{j,k}^{n+\frac{1}{2}}. \tag{6.19}$$

To find $W^{n+\frac{1}{2}}$ we have to solve (M_y-1) systems of (M_x-1) equations and then solve (M_x-1) systems of (M_y-1) equations to compute W^{n+1}. Each of these systems is tridiagonal which makes their solution very fast.

We have to supplement the solutions of equations (6.18-6.19) with boundary conditions. The boundary conditions for (6.19) follow from the Dirichlet boundary condition for the differential problem (6.14). Deriving the boundary conditions for equation (6.18) we should have in mind the accuracy of the scheme as it can be easily destroyed by the wrong choice of boundary data.

Adding the left hand side of (6.19) to the right hand side of (6.18) and solving for $w_{j,k}^{n+\frac{1}{2}}$ we get

$$w_{j,k}^{n+\frac{1}{2}}=\frac{1}{2}\Big(1-\frac{\delta t}{2}a^2\delta_{yy}\Big)w_{j,k}^{n+1}+\frac{1}{2}\Big(1+\frac{\delta t}{2}a^2\delta_{yy}\Big)w_{j,k}^{n}.$$

Thus, for the Peaceman-Rachford scheme with Dirichlet boundary conditions, we should use the following boundary conditions for $w_{j,k}^{n+\frac{1}{2}}$

$$w_{0,k}^{n+\frac{1}{2}}=\frac{1}{2}\Big(1-\frac{\delta t}{2}a^2\delta_{yy}\Big)q(t_{n+1},0,y_k)+\frac{1}{2}\Big(1+\frac{\delta t}{2}a^2\delta_{yy}\Big)q(t_n,0,y_k),$$
$$w_{M_x,k}^{n+\frac{1}{2}}=\frac{1}{2}\Big(1-\frac{\delta t}{2}a^2\delta_{yy}\Big)q(t_{n+1},1,y_k)+\frac{1}{2}\Big(1+\frac{\delta t}{2}a^2\delta_{yy}\Big)q(t_n,1,y_k).$$

The boundary conditions for $w_{j,k}^{n+\frac{1}{2}}$ on the boundaries $y=0$ and $y=1$ are not required as these values can be computed from (6.18).

The truncation error for the Peaceman-Rachford scheme is readily calculated from the unsplit form (6.17). Expanding the terms in equation (6.17) we obtain the Peaceman-Rachford scheme in the form

$$\begin{aligned}\frac{w_{j,k}^{n+1}-w_{j,k}^{n}}{\delta t}=&\frac{1}{2}a^2\delta_{xx}\big(w_{j,k}^{n+1}+w_{j,k}^{n}\big)+\frac{1}{2}a^2\delta_{yy}\big(w_{j,k}^{n+1}+w_{j,k}^{n}\big)\\&-\frac{\delta t}{4}a^4\delta_{xx}\delta_{yy}\big(w_{j,k}^{n+1}-w_{j,k}^{n}\big).\end{aligned}$$

Multiplying this equation by δt and expanding all terms in Taylor's series we deduce that the leading terms are

$$\begin{aligned}\Psi^n(u) \approx & \frac{1}{24}(\delta t)^2\frac{\partial^3 u}{\partial t^3} - \frac{1}{12}(\delta x)^2\frac{\partial^4 u}{\partial x^4} - \frac{1}{12}(\delta y)^2\frac{\partial^4 u}{\partial y^4} \\ & - \frac{1}{8}(\delta t)^2\frac{\partial^4 u}{\partial t^2\partial x^2} - \frac{1}{8}(\delta t)^2\frac{\partial^4 u}{\partial t^2\partial y^2} + \frac{1}{4}(\delta t)^2\frac{\partial^5 u}{\partial t\partial x^2\partial y^2} \\ & = O\big(|\delta t|^2 + |\delta x|^2 + |\delta y|^2\big).\end{aligned}$$

Hence, the Peaceman-Rachford scheme is second-order accurate in δt, δx, and δy similarly to the two-dimensional Crank-Nicolson scheme.

To analyze the scheme stability it is more convenient to use the two step version of the scheme. Applying the discrete Fourier transform to equations (6.18-6.19) we get

$$\begin{aligned}\Big(1 + 2a^2\lambda_x \sin^2\frac{\xi_x}{2}\Big)\hat{W}^{n+\frac{1}{2}} &= \Big(1 - 2a^2\lambda_y \sin^2\frac{\xi_y}{2}\Big)\hat{W}^n, \\ \Big(1 + 2a^2\lambda_y \sin^2\frac{\xi_y}{2}\Big)\hat{W}^{n+1} &= \Big(1 - 2a^2\lambda_x \sin^2\frac{\xi_x}{2}\Big)\hat{W}^{n+\frac{1}{2}}.\end{aligned}$$

From these two equations, we obtain the amplification function

$$G(\xi_x, \xi_y) = \frac{\Big(1 - 2a^2\lambda_x \sin^2\frac{\xi_x}{2}\Big)\Big(1 - 2a^2\lambda_y \sin^2\frac{\xi_y}{2}\Big)}{\Big(1 + 2a^2\lambda_x \sin^2\frac{\xi_x}{2}\Big)\Big(1 + 2a^2\lambda_y \sin^2\frac{\xi_y}{2}\Big)}.$$

From that expression we see that $|G(\xi_x, \xi_y)| \leq 1$. Hence, the Peaceman-Rachford scheme is unconditionally stable. Since we have a consistent, stable scheme, then, by the Lax equivalence theorem, the scheme is convergent.

Douglas-Rachford scheme. The Douglas-Rachford scheme is obtained by splitting the two-dimensional Euler implicit scheme

$$\big(1 - \delta t\, a^2(\delta_{xx} + \delta_{yy})\big)w_{j,k}^{n+1} = w_{j,k}^n. \tag{6.20}$$

To factor the left hand side of this equation into

$$(1 - \delta t\, a^2\delta_{xx})(1 - \delta t\, a^2\delta_{yy})w_{j,k}^{n+1}$$

we have to add to the left hand side of (6.20) the term

$$(\delta t)^2 a^4 \delta_{xx}\delta_{yy} w_{j,k}^{n+1}.$$

To compensate for this term we add to the right hand side

$$(\delta t)^2 a^4 \delta_{xx}\delta_{yy} w^n_{j,k}.$$

We already know from the analysis of the Peaceman-Rachford scheme that the difference

$$(\delta t)^2 a^4 \delta_{xx}\delta_{yy}\big(w^{n+1}_{j,k} - w^n_{j,k}\big) \tag{6.21}$$

is a higher-order term.

Then we obtain the one-step version of the Douglas-Rachford scheme

$$(1-\delta t\, a^2\delta_{xx})(1-\delta t\, a^2\delta_{yy})w^{n+1}_{j,k} = \big(1+(\delta t)^2 a^4\delta_{xx}\delta_{yy}\big)w^n_{j,k}.$$

Since this scheme differs from the Euler implicit scheme by the higher-order term (6.21) the accuracy of the Douglas-Rachford scheme is the same as the Euler implicit scheme, i.e., first-order in δt and second-order in δx and δy.

The splitting form of the Douglas-Rachford scheme is

$$\begin{aligned}(1-\delta t\, a^2\delta_{xx})w^*_{j,k} &= (1+\delta t\, a^2\delta_{yy})w^n_{j,k},\\ (1-\delta t\, a^2\delta_{yy})w^{n+1}_{j,k} &= w^*_{j,k} - \delta t\, a^2\delta_{yy}w^n_{j,k}.\end{aligned} \tag{6.22}$$

As with the Peaceman-Rachford scheme, we have to select carefully the boundary conditions for W^*. From (6.22) we find

$$w^*_{j,k} = (1-\delta t\, a^2\delta_{yy})w^{n+1}_{j,k} + \delta t\, a^2\delta_{yy}w^n_{j,k}.$$

Thus the boundary conditions for W^* at $j=0$ and $j=M_x$ can be derived from the Dirichlet boundary conditions

$$\begin{aligned}w^*_{0,k} &= (1-\delta t\, a^2\delta_{yy})q(t_{n+1},0,y_k) + \delta t\, a^2\delta_{yy}q(t_n,0,y_k),\\ w^*_{M_x,k} &= (1-\delta t\, a^2\delta_{yy})q(t_{n+1},1,y_k) + \delta t\, a^2\delta_{yy}q(t_n,1,y_k).\end{aligned}$$

We proceed to the stability analysis of the Douglas-Rachford scheme using the unsplit version of the scheme. Applying the discrete Fourier transform we get

$$\begin{aligned}&\Big(1+4a^2\lambda_x\sin^2\frac{\xi_x}{2}\Big)\Big(1+4a^2\lambda_y\sin^2\frac{\xi_y}{2}\Big)\hat{W}^{n+1}\\ &\qquad = \Big(1+16a^4\lambda_x\lambda_y\sin^2\frac{\xi_x}{2}\sin^2\frac{\xi_y}{2}\Big)\hat{W}^n.\end{aligned}$$

From the above equation, we obtain the amplification function

$$G(\xi_x,\xi_y) = \frac{\Big(1+16a^4\lambda_x\lambda_y\sin^2\frac{\xi_x}{2}\sin^2\frac{\xi_y}{2}\Big)}{\Big(1+4a^2\lambda_x\sin^2\frac{\xi_x}{2}\Big)\Big(1+4a^2\lambda_y\sin^2\frac{\xi_y}{2}\Big)}.$$

It is easy to see that $0 \le G(\xi_x,\xi_y) \le 1$. Hence, the scheme is unconditionally stable. As a consistent scheme, it is also convergent.

Additional topics

Arbitrary domain

We have investigated two-dimensional finite difference schemes in a unit square to simplify the presentation. Real problems require a domain U complicated in shape. In setting up difference schemes we can use a regular grid and adopt finite differences to the irregular boundary of U. An alternative approach is based on irregular grids better adapted to the boundary. Both of these approaches have some advantages but also specific difficulties. Below, we present the use of regular grids. The analysis of irregular meshes is beyond the scope of these lecture notes.

Consider an open, compact domain $U \subset \mathbb{R}^2$ embedded in a larger rectangular set D_U. A regular mesh J_U covers the whole D_U. The regularity of J_U means that spacing between grid points on any line parallel to the $0x$ axis is constant and equals δx; the same applies to spacing δy along the $0y$ axis. But the regularity does not exclude that $\delta x \neq \delta y$.

The points of J_U, which are inside U, are called the *interior points*. The points on the intersections of mesh lines with the boundary of U are called the *boundary points*. A boundary point can be a point belonging to J_U, but in the majority of cases, it is a point situated between two points in J_U. We call the point $(x_j, y_k) \in J_U$ *regular* if it is an interior point and its distance from any boundary point is not smaller than the grid spacing. The interior points that are closer to some boundary point are called *irregular*. For regular points operators δ_{xx} and δ_{yy} act in the way defined earlier for $U = [0,1]^2$. To define these operators for irregular points take an irregular point (x_j, y_k) and assume that the irregularity holds in the x direction only. Then only δ_{xx} is changed. We have three cases:

1. If (x^*_{j-1}, y_k) is a boundary point, where the star superscript indicates that the distance between x^*_{j-1} and x_j can be smaller than δx, and (x_{j+1}, y_k) is a regular point then

$$\delta_{xx} w^n_{j,k} = \frac{1}{\delta x}\left(\frac{u(t_n, x_{j+1}, y_k) - u(t_n, x_j, y_k)}{\delta x} - \frac{u(t_n, x_j, y_k) - u(t_n, x^*_{j-1}, y_k)}{\delta x^-}\right),$$

   where $\delta x^- = |x_j - x^*_{j-1}|$ and obviously $\delta x^- \leq \delta x$.

2. If (x^*_{j+1}, y_k) is a boundary point and (x_{j-1}, y_k) is a regular point then

$$\delta_{xx} w^n_{j,k} = \frac{1}{\delta x}\left(\frac{u(t_n, x^*_{j+1}, y_k) - u(t_n, x_j, y_k)}{\delta x^+} - \frac{u(t_n, x_j, y_k) - u(t_n, x_{j-1}, y_k)}{\delta x}\right),$$

where $\delta x^+ = |x^*_{j+1} - x_j|$.

3. Both points (x^*_{j+1}, y_k) and (x^*_{j-1}, y_k) are boundary points then

$$\delta_{xx} w^n_{j,k} = \frac{1}{\delta x}\left(\frac{u(t_n, x^*_{j+1}, y_k) - u(t_n, x_j, y_k)}{\delta x^+} - \frac{u(t_n, x_j, y_k) - u(t_n, x^*_{j-1}, y_k)}{\delta x^-}\right).$$

Similar modifications of δ_{yy} are required when (x_j, y_k) is irregular in the y direction.

These modifications make the implementation of ADI algorithms more complicated. But the essential advantage of these schemes that we only solve linear systems with tridiagonal matrices remains valid.

Mixed derivatives

We extend the considerations to the anisotropic diffusion equation with a mixed second-order derivative

$$\frac{\partial u}{\partial t} - a_{11}\frac{\partial^2 u}{\partial x^2} - 2a_{12}\frac{\partial^2 u}{\partial x \partial y} - a_{22}\frac{\partial^2 u}{\partial y^2} = 0 \tag{6.23}$$

defined on $U = [0,1]^2$ with Dirichlet boundary conditions.

The difference schemes we have used for the heat equation will now involve nine points on each time level to approximate the spatial derivatives. That cause much greater difficulty compared to the five-point schemes used for the heat equation. Besides, the use of the ellipticity condition $a_{11}a_{22} > a_{12}^2$ to control the growth of the terms coming from the approximation of mixed derivatives can be insufficient, particularly, when the coefficients a_{11} and a_{22} are of different magnitude. Hence, numerical schemes for such equations require careful design to obtain accurate and convergent algorithms.

Fortunately, in ADI schemes we can put the approximation of mixed derivatives into the explicit part of the algorithm. We confine the discussion to the Peaceman-Rachford scheme. For equation (6.23) we have the following extension of the Peaceman-Rachford scheme

$$\Big(1 - \frac{\delta t}{2} a_{11}\delta_{xx}\Big) w^{n+\frac{1}{2}}_{j,k} = \Big(1 + \frac{\delta t}{2} a_{22}\delta_{yy} + \delta t\, a_{12}\delta_x\delta_y\Big) w^n_{j,k},$$
$$\Big(1 - \frac{\delta t}{2} a_{22}\delta_{yy}\Big) w^{n+1}_{j,k} = \Big(1 + \frac{\delta t}{2} a_{11}\delta_{xx} + \delta t\, a_{12}\delta_x\delta_y\Big) w^{n+\frac{1}{2}}_{j,k}.$$

Unfortunately, this scheme is only first-order accurate in time. The scheme can be upgraded to a scheme that is second-order accurate in time by introducing additional intermediate values

$$\widetilde{W}^n = \frac{3}{2}W^n - \frac{1}{2}W^{n-1}.$$

The improved scheme is as follows

$$\Big(1 - \frac{\delta t}{2}a_{11}\delta_{xx}\Big)w_{j,k}^{n+\frac{1}{2}} = \Big(1 + \frac{\delta t}{2}a_{22}\delta_{yy}\Big)w_{j,k}^{n} + \delta t\, a_{12}\delta_x\delta_y\widetilde{w}_{j,k}^{n},$$
$$\Big(1 - \frac{\delta t}{2}a_{22}\delta_{yy}\Big)w_{j,k}^{n+1} = \Big(1 + \frac{\delta t}{2}a_{11}\delta_{xx}\Big)w_{j,k}^{n+\frac{1}{2}} + \delta t\, a_{12}\delta_x\delta_y\widetilde{w}_{j,k}^{n}.$$

The boundary conditions for this scheme can be obtained similarly like for the original Peaceman-Rachford scheme

$$w_{0,k}^{n+\frac{1}{2}} = \frac{1}{2}\Big(1 - \frac{\delta t}{2}a_{22}\delta_{yy}\Big)q(t_{n+1}, 0, y_k) + \frac{1}{2}\Big(1 + \frac{\delta t}{2}a_{22}\delta_{yy}\Big)q(t_n, 0, y_k),$$
$$w_{M_x,k}^{n+\frac{1}{2}} = \frac{1}{2}\Big(1 - \frac{\delta t}{2}a_{22}\delta_{yy}\Big)q(t_{n+1}, 1, y_k) + \frac{1}{2}\Big(1 + \frac{\delta t}{2}a_{22}\delta_{yy}\Big)q(t_n, 1, y_k).$$

Lower order terms

There is no doubt about where to place first-order derivatives in ADI schemes. The first-order derivative with respect to x should go with the second-order x derivative. The same applies to the derivatives with respect to y. As we know from Section 6.4, the way we approximate the first-order derivatives influences greatly the scheme stability. The results obtained in that section extend straightforwardly to ADI schemes. The formulas used for the approximation of first-order derivatives and the placing of the corresponding terms in the implicit or explicit part of the scheme decide on the accuracy of the scheme. These effects can be easily investigated by adapting the computations of Section 6.4.

For non-homogeneous equations with a non-homogeneous term $f(t, x, y)$ it is not always clear where such a term should go in an ADI scheme. Using a wrong approximation, we can destroy the scheme's accuracy. The suggested solution is to start with an integral form of the differential equation (in an analogy with equations of mathematical physics, that integral form is called the conservation law formulation). Then approximate the integral of $f(t, x, y)$ in a way that preserves the desired order of accuracy. We are not going into details about these computations. To illustrate that the good placement of the nonhomogeneous term is not obvious, we write the Peaceman-Rachford scheme for the equation

$$\frac{\partial u}{\partial t} - a^2\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = f(t, x, y).$$

The Peaceman-Rachford scheme which preserves the second-order accuracy is

$$\Big(1-\frac{\delta t}{2}a^2\delta_{xx}\Big)w_{j,k}^{n+\frac{1}{2}}=\Big(1+\frac{\delta t}{2}a^2\delta_{yy}\Big)w_{j,k}^{n}+\frac{\delta t}{2}f_{j,k}^{n},$$
$$\Big(1-\frac{\delta t}{2}a^2\delta_{yy}\Big)w_{j,k}^{n+1}=\Big(1+\frac{\delta t}{2}a^2\delta_{xx}\Big)w_{j,k}^{n+\frac{1}{2}}+\frac{\delta t}{2}f_{j,k}^{n+1}.$$

Three dimensional schemes

Below we give a brief presentation of three-dimensional ADI schemes. As before, we consider the heat equation

$$\frac{\partial u}{\partial t}-a^2\bigg(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}+\frac{\partial^2 u}{\partial z^2}\bigg)=0.$$

The extension of two-dimensional ADI schemes to three dimensions in not straightforward as the example of the Peaceman-Rachford scheme shows. The three-dimensional Peaceman-Rachford scheme is not unconditionally stable and is only first-order accurate in time. On the other hand, the three-dimensional Douglas-Rachford scheme

$$(1-\delta t\,a^2\delta_{xx})w_{j,k,l}^{*}=\big(1+\delta t\,a^2(\delta_{yy}+\delta_{zz})\big)w_{j,k,l}^{n},$$
$$(1-\delta t\,a^2\delta_{yy})w_{j,k,l}^{**}=w_{j,k,l}^{*}-\delta t\,a^2\delta_{yy}w_{j,k,l}^{n},$$
$$(1-\delta t\,a^2\delta_{zz})w_{j,k,l}^{n+1}=w_{j,k,l}^{**}-\delta t\,a^2\delta_{zz}w_{j,k,l}^{n}$$

is unconditionally stable and $O\big(|\delta t|+|\delta x|^2+|\delta y|^2+|\delta z|^2\big)$ accurate, exactly like in two dimensions.

To obtain unconditionally stable ADI schemes, we have to proceed like in two dimensions, i.e., start from a stable scheme, like the Crank-Nicolson or the implicit Euler scheme, and perform an appropriate splitting. We will illustrate such splitting for the three-dimensional Crank-Nicolson scheme

$$\Big(1-\frac{\delta t}{2}a^2(\delta_{xx}+\delta_{yy}+\delta_{zz})\Big)w_{j,k,l}^{n+1}=\Big(1+\frac{\delta t}{2}a^2(\delta_{xx}+\delta_{yy}+\delta_{zz})\Big)w_{j,k,l}^{n}\tag{6.24}$$

which is unconditionally stable and $O\big(|\delta t|^2+|\delta x|^2+|\delta y|^2+|\delta z|^2\big)$ accurate. Adding to the left hand side the expression

$$\frac{(\delta t)^2}{4}a^4(\delta_{xx}\delta_{yy}+\delta_{xx}\delta_{zz}+\delta_{yy}\delta_{zz})w_{j,k,l}^{n+1}-\frac{(\delta t)^3}{8}a^6\delta_{xx}\delta_{yy}\delta_{zz}w_{j,k,l}^{n+1}$$

and the similar expression for $w^n_{j,k,l}$ to the right hand side, we obtain the scheme

$$\begin{aligned}\Big(1-\frac{\delta t}{2}a^2\delta_{xx}\Big)\Big(1-\frac{\delta t}{2}a^2\delta_{yy}\Big)\Big(1-\frac{\delta t}{2}a^2\delta_{zz}\Big)w^{n+1}_{j,k,l}\\ =\Big(1+\frac{\delta t}{2}a^2\delta_{xx}\Big)\Big(1+\frac{\delta t}{2}a^2\delta_{yy}\Big)\Big(1+\frac{\delta t}{2}a^2\delta_{zz}\Big)w^{n}_{j,k,l}.\end{aligned} \tag{6.25}$$

Collecting the added terms

$$\begin{aligned}\frac{(\delta t)^2}{4}a^4(\delta_{xx}\delta_{yy}+\delta_{xx}\delta_{zz}+\delta_{yy}\delta_{zz})\big(w^{n+1}_{j,k,l}-w^n_{j,k,l}\big)\\ -\frac{(\delta t)^3}{8}a^6\delta_{xx}\delta_{yy}\delta_{zz}\big(w^{n+1}_{j,k,l}+w^n_{j,k,l}\big),\end{aligned}$$

we obtain an expression of order $O\big(|\delta t|^3\big)$. Hence, up to $O\big(|\delta t|^2\big)$ the schemes (6.24) and (6.25) are equivalent.

We now write (6.25) in the form

$$\begin{aligned}\Big(1-\frac{\delta t}{2}a^2\delta_{xx}\Big)\Big(1-\frac{\delta t}{2}a^2\delta_{yy}\Big)\Big(1-\frac{\delta t}{2}a^2\delta_{zz}\Big)\big(w^{n+1}_{j,k,l}-w^n_{j,k,l}\big)\\ =\delta t\,a^2(\delta_{xx}+\delta_{yy}+\delta_{zz})w^n_{j,k,l}+\frac{(\delta t)^3}{4}a^6\delta_{xx}\delta_{yy}\delta_{zz}w^n_{j,k,l}.\end{aligned} \tag{6.26}$$

Since the last term in equation (6.26) is of order $O\big(|\delta t|^3\big)$, it can be dropped without destroying the scheme accuracy. We split the modified scheme using the so-called Δ-formulation. Then we arrive at the Douglas-Gunn scheme

$$\begin{aligned}&(1-\frac{\delta t}{2}a^2\delta_{xx})\Delta w^*_{j,k,l}=\delta t\,a^2(\delta_{xx}+\delta_{yy}+\delta_{zz})w^n_{j,k,l},\\ &(1-\frac{\delta t}{2}a^2\delta_{yy})\Delta w^{**}_{j,k,l}=\Delta w^*_{j,k,l},\\ &(1-\frac{\delta t}{2}a^2\delta_{zz})\Delta w_{j,k,l}=\Delta w^{**}_{j,k,l},\\ &\Delta w_{j,k,l}=w^{n+1}_{j,k,l}-w^n_{j,k,l}.\end{aligned}$$

The scheme is $O\big(|\delta t|^2+|\delta x|^2+|\delta y|^2+|\delta z|^2\big)$ accurate as the Crank-Nicolson

scheme. Applying the discrete Fourier transform to this scheme we find

$$\begin{aligned}
&\Big(1+2a^2\lambda_x \sin^2\frac{\xi_x}{2}\Big)\Big(1+2a^2\lambda_y \sin^2\frac{\xi_y}{2}\Big)\Big(1+2a^2\lambda_z \sin^2\frac{\xi_z}{2}\Big)\hat{W}^{n+1}\\
&\quad=\Big(1-2a^2\lambda_x \sin^2\frac{\xi_x}{2}-2a^2\lambda_y \sin^2\frac{\xi_y}{2}-2a^2\lambda_z \sin^2\frac{\xi_z}{2}\\
&\qquad+4a^4\lambda_x\lambda_y \sin^2\frac{\xi_x}{2}\sin^2\frac{\xi_y}{2}+4a^4\lambda_x\lambda_z \sin^2\frac{\xi_x}{2}\sin^2\frac{\xi_z}{2}\\
&\qquad+4a^4\lambda_y\lambda_z \sin^2\frac{\xi_y}{2}\sin^2\frac{\xi_z}{2}+8a^6\lambda_x\lambda_y\lambda_z \sin^2\frac{\xi_x}{2}\sin^2\frac{\xi_y}{2}\sin^2\frac{\xi_z}{2}\Big)\hat{W}^n.
\end{aligned}$$

It is easy to check that the amplification function $G(\xi_x, \xi_y, \xi_z)$ which follows from this expression fulfills the condition

$$-1 \leq G(\xi_x, \xi_y, \xi_z) \leq 1.$$

Hence, the scheme is unconditionally stable. Being stable and consistent the scheme is convergent.

Chapter 7

Finite element methods

The finite difference methods with uniform grids are easy to implement in computer codes. But these methods are not flexible enough for complex domains and low smoothness data. An alternative approach is based on the Galerkin approximation described in Section 5.3 and is called the *finite element method*.

The finite element method has become the most important method for approximating the solutions of partial differential equations, in particular of elliptic and parabolic types. The method is based on the variational form of boundary value problems and approximates the exact solution by a piecewise polynomial function. In that approach, one can easily handle complicated domains and solutions of minimal regularity. It permits an accurate error analysis allowing to estimate the cost of the made approximation. The method results in a finite algebraic system of equations for the approximate solution. But, unlike the finite difference method, the approximate solution is known in the whole domain as a piecewise polynomial function and not just as a set of values in grid points.

7.1 Finite elements for elliptic equations

The Galerkin method applied in Section 5.3 to second-order parabolic equations can also be applied to the elliptic equation

$$\begin{aligned} &\mathcal{A}u = f, \quad \text{in } U, \\ &u|_{\partial U} = 0, \end{aligned} \tag{7.1}$$

where $\mathcal{A}$ is the uniformly elliptic operator in divergence form given by (5.7) with coefficients a_{ij}, b_i, and c sufficiently smooth.

Before implementing the Galerkin approximation we modify slightly the elliptic problem (7.1). Let us recall that in the proof of existence of weak solutions in

Section 5.2 we have modified the elliptic problem (7.1) to the following

$$\begin{aligned} &\mathcal{A}u + \mu u = f, \quad \text{in } U, \\ &u|_{\partial U} = 0, \end{aligned}$$

and have proved that for a sufficiently large μ this problem possesses a unique solution. That corresponds to adding to the function $c(x)$ in the definition of $\mathcal{A}$ a large constant μ. To simplify the presentation we assume now that a sufficiently large constant has been already added to $c(x)$ and the bilinear form $B[u,v]$ (see Theorem 5.20) is given by the expression

$$B[u,v] = \int_U \Big(\sum_{i,j=1}^{d} a_{ij}(x) \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} + \sum_{i=1}^{d} b_i(x) \frac{\partial u}{\partial x_i} v + \bar{c}(x) uv \Big) dx, \tag{7.2}$$

where $\bar{c}(x) = c(x) = \mu$. In what follows we will skip the bar over $c(x)$ assuming that the function $c(x)$ in the definition of $\mathcal{A}$ is such that estimate (7.5) (see below) holds.

Hence we will consider the weak formulation of the Dirichlet problem (7.1)

$$\text{find } u \in H_0^1(U): \; B[u,v] = (f,v), \quad \forall v \in H_0^1(U). \tag{7.3}$$

The following estimates can be derived from the estimates of Theorem 5.20

$$|B[u,v]| \le \alpha \|u\|_{H_0^1(U)} \|v\|_{H_0^1(U)}, \tag{7.4}$$

$$\beta \|u\|^2_{H_0^1(U)} \le B[u,u]. \tag{7.5}$$

The bilinear form $B[u,v]$ is said to be *coercive* when it fulfills inequality (7.5).

To implement the Galerkin approximation, we have to select a basis $\{\psi_k\}_{k=1}^{\infty}$ and define the functions

$$u_m = \sum_{k=1}^{m} d_m^k \psi_k,$$

to approximate the Dirichlet problem (7.1) by the weak formulation

$$B[u_m, \psi_k] = (f, \psi_k), k = 1, \ldots, m. \tag{7.6}$$

Using the definition of u_m, we can reduce (7.6) to the linear system

$$\sum_{i=1}^{m} e^{ki} d_m^i = f_k, k = 1, \ldots, m, \tag{7.7}$$

where $e^{ki} = B[\psi_i, \psi_k]$ and $f_k = (f, \psi_k)$. This system of linear equations possesses a unique solution since $\{e^{ki}\}_{k,i=1}^m$ is a nonsingular matrix due to the coerciveness of B.

In Section 5.3 we have chosen as $\{\psi_k\}_{k=1}^\infty$ the eigenvectors of operator $\mathcal{A}_0 = \sum_{i,j=1}^d -\frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial u}{\partial x_j}\right)$ which form an orthonormal basis in $L^2(U)$ and an orthogonal basis in $H_0^1(U)$. It appears, however, that using the eigenvectors of $\mathcal{A}_0$ as the basis is not efficient for numerical implementation. The construction of the basis suitable for numerical computations is the most challenging problem in the implementation of the Galerkin approximation. One of the possibilities is to use the so-called *finite elements*. First, we will present their construction in one dimension.

The finite element method in the one-dimensional case. Let U be a finite interval. Without loss of generality, we can take $U = (0, 1)$ and

$$\mathcal{A}u(x) := -\tfrac{d}{dx}(a^2(x)\tfrac{du(x)}{dx}) + b(x)\tfrac{du(x)}{dx} + c(x)u(x) = f(x), \quad \text{in } U,$$

with $u(0) = u(1) = 0$.

We define a partition $\mathcal{T}_h$ dividing $(0, 1)$ into M sub-intervals $K_j = (x_{j-1}, x_j)$, called *elements*, with width $h_j = x_j - x_{j-1}$ such that

$$0 = x_0 < x_1 < \cdots < x_{M-1} < x_M = 1.$$

We set $h = \max_j h_j$ and for $h \in H_I \subset (0, +\infty)$ construct the family of spaces

$$X_h^r = \{v_h \in C(\bar{U}):\ v_h\big|_{K_j} \in \Pi^r, \quad \forall K_j \in \mathcal{T}_h\}, r = 1, 2, \ldots, \tag{7.8}$$

where $\Pi^r = \Pi^r(K)$ denotes the space of polynomials of degree not greater than r defined in K.

We replace $H_0^1(U)$, the functional space for the Galerkin approximation, by the finite dimensional space $V_h = X_h^r \cap H_0^1(U)$ with a fixed r. We construct now a basis $\{\phi_i\}$ in V_h. Let us begin with X_h^1. X_h^1 is a space of piecewise linear functions. Each such function is uniquely determined by its values in the vertices x_j. Since we have $M + 1$ vertices then defining $M + 1$ basis functions ϕ_i, $i = 0, \ldots, M$, we define the space X_h^1. The basis which is mostly used is the basis consisting of functions ϕ_i such that

$$\phi_i(x_j) = \delta_{ij}, \quad i, j = 0, 1, \ldots, M, \tag{7.9}$$

where δ_{ij} is the Kronecker delta.

These functions, called the *"hat" functions* due to their shape, are given by the expression

$$\phi_i(x) = \begin{cases} \frac{x-x_{i-1}}{x_i-x_{i-1}}, & \text{for } x_{i-1} \le x \le x_i, \\ \frac{x_{i+1}-x}{x_i-x_{i-1}}, & \text{for } x_i \le x \le x_{i+1}, \\ 0, & \text{otherwise.} \end{cases} \tag{7.10}$$

We now define X_h^2, a space of piecewise quadratic functions. Since a quadratic function is uniquely defined by its values in 3 distinct points, to define a basis of quadratic functions in each sub-interval, we have to supplement vertices x_j by midpoints of each sub-interval (x_{j-1}, x_j) obtaining points

$$0 = x_0 < x_1 < \cdots < x_{2M} = 1.$$

Then $2M + 1$ basis functions ϕ_i of X_h^2 can be such that

$$\phi_i(x_j) = \delta_{ij}, \quad i, j = 0, 1, \ldots, 2M.$$

That construction can be extended to arbitrary integers $r > 0$.

We now return to the one-dimensional elliptic problem

$$\begin{aligned} &- \tfrac{d}{dx}(a^2(x)\tfrac{du(x)}{dx}) + b(x)\tfrac{du(x)}{dx} + c(x)u(x) = f(x), \quad 0 < x < 1, \\ &u(0) = 0 = u(1). \end{aligned} \tag{7.11}$$

Our goal is to find an approximate solution to this problem in the family of spaces V_h with $h \in H_I$

$$V_h = \{v_h \in X_h^1 \colon\ v_h(0) = 0 = v_h(1)\}.$$

These spaces are finite-dimensional subspaces of $H_0^1(0, 1)$ since functions from X_h^1 are differentiable on $(0, 1)$ except a finite number of vertices x_j.

We construct the Galerkin approximation by finding $u_h \in V_h$ which for each $v_h \in V_h$ solves the equation

$$\begin{aligned} B_I[u_h, v_h] &:= \int_0^1 \Big(a^2(x)\tfrac{du_h}{dx}\tfrac{dv_h}{dx} + b(x)\tfrac{du_h}{dx}v_h + c(x)u_h v_h \Big) dx \\ &= \int_0^1 f(x)v_h dx. \end{aligned} \tag{7.12}$$

Using the basis of the hat functions in X_h^1 we can expand $u_h(x) = \sum_{i=1}^{M-1} w_i\phi_i(x)$. Then taking as v_h basis functions $\phi_j(x)$ we can rewrite (7.12) as the linear system

$$A_h W = F, \tag{7.13}$$

where A_h is the so-called *stiffness matrix* with elements $a_{ij} = B_I[\phi_i, \phi_j]$, $W = (w_1, \dots w_{M-1})$ is the vector of components of u_h in the basis ϕ_i, and F is the *load vector* with components $f_i = (f, \phi_i)$ (lack of indices $i = 0$ and $i = M$ is due to the boundary conditions $u_h(0) = u_h(1) = 0$).

According to our assumptions $B_I[u, v]$ is coercive, therefore A_h is nonsingular and there is a unique solution of system (7.13). Then $u_h(x) = \sum_{i=1}^{M-1} w_i \phi_i(x)$ is an approximate weak solution of the Dirichlet problem (7.11). Our ultimate goal is to estimate the error of this approximation.

Before we analyze the error, we define an interpolation operator in X_h^1 and discuss its properties. For $v \in C([0,1])$ we define the *interpolant* $I_h^1 v$ determined by the partition $\mathcal{T}_h$. For each node x_i, $i = 0, \dots, M$, we set

$$I_h^1 v(x_i) = v(x_i).$$

Using the basis $\{\phi_i\}_{i=0}^M$ of X_h^1 we can write the interpolant in the following way

$$I_h^1 v(x) = \sum_{i=0}^{M} v(x_i)\phi_i(x).$$

The operator $I_h^1 \colon\ C([0,1]) \to X_h^1$ is called the *interpolation operator*. Let us observe that in one dimension functions from $H^1(0,1)$ are continuous, hence the interpolation operator is also well defined in $H^1(0,1)$.

Analogously, we can define the interpolation operator $I_h^r \colon\ C([0,1]) \to X_h^r$ for $r > 1$. First, we interpolate v on each $K_j \in \mathcal{T}_h$ projecting v on $\Pi^r(K_j)$. This projection is defined, similarly like in the case of X_h^1, on the basis functions $\phi_{i,j} \in \Pi^r(K_j)$ such that

$$\phi_{i,j}(x_{j,l}) = \delta_{il},$$

where $x_{j,l}$, $l = 0, \dots, r$ are $r+1$ nodes of K_j. Denoting this interpolation as $I_{K_j}^r v$ we can write

$$I_h^r v\Big|_{K_j} = I_{K_j}^r\big(v\big|_{K_j}\big) \quad \forall K_j \in \mathcal{T}_h. \tag{7.14}$$

Then we have the following theorem.

THEOREM. 7.1 *Let $v \in H^{r+1}(0,1)$, for $r \geq 1$, and $I_h^r v$ be its interpolant in X_h^r defined by* (7.14)*. Then the following estimate holds*

$$|v - I_h^r v|_{H^m(0,1)} \leq C_{m,r} h^{r+1-m} |v|_{H^{r+1}(0,1)}, \quad m = 0, 1,$$

where the constant $C_{m,r}$ is independent of v and h.

Here $|\cdot|_{H^k(U)}$ denotes the seminorm in $H^k(U)$ defined by the derivatives of order k, i.e.,

$$|u|_{H^k(U)} = \left(\sum_{|\alpha|=k} \int_U \left|D^\alpha u\right|^2 dx \right)^{\frac{1}{2}}.$$

Proof. We prove the theorem only for $r = 1$. Due to the Sobolev inequality if $v \in H^2(0,1)$ then $v \in C^1(0,1)$ and $I_h^1 v$ is differentiable except a finite number of nodes x_j. Let $z = v - I_h^1 v$. Since $z(x_i) = 0$, $i = 0, \ldots, M$, then by Rolle's theorem in each $K_j = (x_{j-1}, x_j)$ there exists $\xi_j \in K_j$ such that $z'(\xi_j) = 0$.

Since $I_h^1 v$ is a linear function in each K_j, then

$$z'(x) = \int_{\xi_j}^{x} z''(s)ds = \int_{\xi_j}^{x} v''(s)ds, \quad \text{for } x \in K_j,$$

where the last integral is well defined since $v \in H^2(0,1)$.

By the Hölder inequality we obtain

$$\begin{aligned} |z'(x)| \le \int_{x_{j-1}}^{x_j} |v''(s)|ds &\le \left(|x_j - x_{j-1}|\right)^{\frac{1}{2}} \left(\int_{x_{j-1}}^{x_j} |v''(s)|^2 ds\right)^{\frac{1}{2}} \\ &\le h^{\frac{1}{2}} \left(\int_{x_{j-1}}^{x_j} |v''(s)|^2 ds\right)^{\frac{1}{2}}. \end{aligned} \tag{7.15}$$

Hence

$$\int_{x_{j-1}}^{x_j} |z'(s)|^2 ds \le h^2 \int_{x_{j-1}}^{x_j} |v''(s)|^2 ds.$$

Since for $x \in K_j$

$$z(x) = \int_{x_{j-1}}^{x} z'(s)ds$$

then, using estimate (7.15), we obtain

$$|z(x)| \le \int_{x_{j-1}}^{x_j} |z'(s)|ds \le h^{\frac{3}{2}} \left(\int_{x_{j-1}}^{x_j} |v''(s)|^2 ds\right)^{\frac{1}{2}}.$$

That gives

$$\int_{x_{j-1}}^{x_j} |z(s)|^2 ds \le h^4 \int_{x_{j-1}}^{x_j} |v''(s)|^2 ds.$$

Summing over all K_j, $j = 1, \ldots, M$, we obtain

$$\begin{aligned} \int_0^1 |z'(s)|^2 ds &\le h^2 \int_0^1 |v''(s)|^2 ds, \\ \int_0^1 |z(s)|^2 ds &\le h^4 \int_0^1 |v''(s)|^2 ds. \end{aligned}$$

These estimates prove the theorem for $r = 1$ with $C_{m,1} = 1$, $m = 0, 1$. ∎

The following lemma is valid for multi-dimensional finite element approximations, which we will discuss in the next subsection.

THEOREM. 7.2 (Céa's lemma) *Let us consider $u \in H_0^1(U)$ a weak solution of the Dirichlet problem* (7.1) *and its Galerkin approximation $u_h \in V_h$ which solves* (7.26) *where U is an open, bounded domain in $\mathbb{R}^d$. Then*

$$\|u - u_h\|_{H_0^1(U)} \leq \frac{\alpha}{\beta} \inf_{v_h \in V_h} \|u - v_h\|_{H_0^1(U)}.$$

Proof. The proof follows from the properties of the bilinear form $B[u, v]$. Let $u \in H_0^1(U)$ solve

$$B[u, v] = (f, v), \quad \forall v \in H_0^1(U),$$

and $u_h \in V_h$ solve

$$B[u_h, v] = (f, v), \quad \forall v \in V_h.$$

Since $V_h \subset H_0^1(U)$ then taking both equalities with $v \in V_h$ we obtain the orthogonality relation

$$B[u - u_h, v] = 0, \quad \forall v \in V_h. \tag{7.16}$$

Let $v_h \in V_h$. Taking $v = u_h - v_h$ in (7.16) we have

$$B[u - u_h, u - u_h] = B[u - u_h, u - v_h] \leq \alpha \|u - u_h\|_{H_0^1(U)} \|u - v_h\|_{H_0^1(U)}.$$

On the other hand, the coerciveness of B gives

$$\beta \|u - u_h\|^2_{H_0^1(U)} \leq B[u - u_h, u - u_h].$$

By the above estimates, we have

$$\|u - u_h\|_{H_0^1(U)} \leq \frac{\alpha}{\beta} \|u - v_h\|_{H_0^1(U)}.$$

Since v_h is an arbitrary vector in V_h, this proves the assertion of the theorem. ■

A simple corollary from the Céa lemma is the following H^1 error estimate of the Galerkin approximation.

THEOREM. 7.3 *Let $u \in H_0^1(0,1)$ be a solution of problem* (7.11) *and u_h a solution of its finite element approximation* (7.12) *in $V_h = H_0^1(0,1) \cap X_h^r$. In addition, let $u \in H^{m+1}(0,1)$ for $m \geq r$. Then the following* a priori *estimate holds*

$$\|u - u_h\|_{H_0^1(0,1)} \leq \frac{\alpha}{\beta} C h^r |u|_{H^{r+1}(0,1)},$$

where the constant C is independent of u and h.

Proof. Inserting in Céa's lemma $v_h = I_h^r u$, we get

$$\|u - u_h\|_{H_0^1(0,1)} \leq \frac{\alpha}{\beta} \|u - I_h^r u\|_{H_0^1(0,1)},$$

By the estimate of Theorem 7.1, we obtain the claim of the theorem. ■

We can obtain a better error estimate in $L^2(0,1)$.

THEOREM. 7.4 *Let $u \in H_0^1(0,1)$ be a solution of problem* (7.11) *and u_h a solution of its finite element approximation* (7.12) *in $V_h = H_0^1(0,1) \cap X_h^r$. In addition, let $u \in H^{m+1}(0,1)$ for $m \geq r$. Then the following* a priori *estimate holds*

$$\|u - u_h\|_{L^2(0,1)} \leq C\, h^{r+1} |u|_{H^{r+1}(0,1)},$$

where the constant C is independent of u and h.

Proof. For an arbitrary $g \in L^2(0,1)$, we consider the following dual problem

$$\text{find } z \in H_0^1(0,1):\ B_I[v,z] = (v,g), \quad \forall v \in H_0^1(0,1).$$

It can be shown (similarly like for problem (7.11)) that the dual problem has a unique solution $z = z(g)$ and the following regularity result holds

$$\|z\|_{H^2(0,1)} \leq C \|g\|_{L^2(0,1)}, \quad \forall g \in L^2(0,1), \tag{7.17}$$

(see the regularity result of Theorem 5.22).

Let $e_h = u - u_h$. Choosing $g = e_h$ and $v = e_h$ we obtain

$$B_I[e_h, z(e_h)] = \|e_h\|^2_{L^2(0,1)}.$$

Due to the Galerkin orthogonality property (7.16), we have

$$B_I[e_h, z] = B_I[u - u_h, z] = B_I[u - u_h, z - I_h^1 z] = B_I[e_h, z - I_h^1 z].$$

Then, using the estimate of Theorem 7.1 and estimate (7.17), we have

$$\begin{aligned} \|e_h\|^2_{L^2(0,1)} &= B_I[e_h, z] = B_I[e_h, z - I_h^1 z] \\ &\leq C \|e_h\|_{H^1(0,1)} \|z - I_h^1 z\|_{H^1(0,1)} \leq C\, h \|e_h\|_{H^1(0,1)} \|z\|_{H^2(0,1)} \\ &\leq C\, h \|e_h\|_{H^1(0,1)} \|e_h\|_{L^2(0,1)}. \end{aligned}$$

Thus we obtain

$$\|e_h\|_{L^2(0,1)} \leq C\, h \|e_h\|_{H^1(0,1)}$$

which together with Theorem 7.3 gives the desired estimate. ■

The multi-dimensional case. To extend the finite element method to multi-dimensional domains we have to define finite elements for multi-dimensional boundary-value problems. Let $U \subset \mathbb{R}^d$ be an open, bounded, convex domain with a Lipschitz boundary. Our first step is to define a triangulation of U. Contrary to the one-dimensional case we now have many possibilities of triangulations. The most popular is to split U into simplices. If $\{x_0, \ldots, x_d\}$ are $d+1$ points in $\mathbb{R}^d$ then the convex hull of $\{x_0, \ldots, x_d\}$ is called a *simplex* when vectors $\{x_1 - x_0, \ldots, x_d - x_0\}$ are linearly independent (in other words $\{x_0, \ldots, x_d\}$ do not belong to a single hyperplane in $\mathbb{R}^d$). Points $\{x_0, \ldots, x_d\}$ are then called the *vertices of the simplex*. The standard simplex in $\mathbb{R}^d$ is the set

$$\hat{K}_d = \{x \in \mathbb{R}^d \colon\ x_i \geq 0, \sum_{i=1}^{d} x_i \leq 1\}. \tag{7.18}$$

This is a unit interval in $\mathbb{R}^1$, a unit triangle in $\mathbb{R}^2$, and a unit tetrahedron in $\mathbb{R}^3$.

The triangulation $\mathcal{T}_h$ of $U \subset \mathbb{R}^d$ is for a given h a division of U into non-overlapping simplices in such a way that the intersection of any two simplices of $\mathcal{T}_h$ is either empty or a common face, where a face of dimension $m < d$ of a simplex is the convex hull of a subset of $m+1$ its vertices. For a simplex $K \in \mathcal{T}_h$, we define its diameter $h_K = diam(K)$. Then the diameter of triangulation $\mathcal{T}_h$ is $h = \max_{K \in \mathcal{T}_h} h_K$.

In general, for a convex domain U with a Lipschitz boundary the sum of all simplices of a triangulation

$$U_h = \operatorname{int}\Big(\bigcup_{K \in \mathcal{T}_h} K \Big)$$

is a proper subset of U. That leads to additional difficulties with an error estimate on $U \setminus U_h$. In what follows, we omit that problem assuming that the shape of U is such that $U = U_h$.

DEFINITION. 7.5 *For each simplex $K \in \mathcal{T}_h$ we define a finite element as a triple $(K, \Pi^r(K), \Sigma)$ where*

(i) K is a closed simplex with a Lipschitz boundary;

(ii) $\Pi^r(K)$ is a space of polynomials of degree not greater than r, dimension $M_r = \binom{r+d}{r}$, and a basis $\{\phi_i\}_{i=1}^{M_r}$;

(iii) Σ is a set of linear functionals called the degrees of freedom *$\sigma_i \colon\ \Pi^r(K) \to \mathbb{R}$, $i = 1, \ldots, M_r$, which are linearly independent.*

The basis of $\Pi^r(K)$ *is selected to be dual to* Σ*, i.e.,* $\sigma_i(\phi_j) = \delta_{ij}$.

The space $\Pi^r(K)$ *is called* unisolvent *with respect to the functionals* $\{\sigma_i\}_{i=1}^{M_r}$ *if for each* $a = (\alpha_1, \ldots, \alpha_{M_r}) \in \mathbb{R}^{M_r}$ *there is exactly one polynomial* $P \in \Pi^r(K)$ *such that* $\sigma_i(P) = \alpha_i$, $i = 1, \ldots, M_r$. *When the space* $\Pi^r(K)$ *is unisolvent, we say also that the functionals* $\{\sigma_i\}_{i=1}^{M_r}$ *are unisolvent.*

DEFINITION. 7.6 *Let* $\Pi^r(K)$ *be unisolvent. There exists in* K *a set of points* $\{a_i\}_{i=1}^{M_r}$ *such that for* $\forall P \in \Pi^r(K)$ $\sigma_i(P) = P(a_i)$, $i = 1, \ldots, M_r$. *These points are called* nodes.

The finite elements whose linear functionals are defined by evaluations on the nodes in K *are called the* Lagrangian finite elements. *For the Lagrangian finite elements we have for every* $P \in \Pi^r(K)$ *the expansion* $P(x) = \sum_{i=1}^{M_r} \sigma_i(P)\phi_i(x)$.

For the Lagrangian finite elements, the functionals σ_i *are often identified with the nodes* a_i*, and then* a_i *are called the degrees of freedom.*

DEFINITION. 7.7 *A family of triangulations* $\mathcal{T}_h$ *parametrized by* $h \in H_I \subset (0, \infty)$ *forms regular triangulations if*

i) 0 *is a limit point of* H_I*;*

ii) there is a constant $C \geq 1$ *such that*

$$h_K \leq C\,\rho(K), \quad \forall K \in \mathcal{T}_h,$$

where $\rho(K)$ *is the radius of the greatest circle inscribed in* K.

DEFINITION. 7.8 *Consider a finite element* $(\hat{K}, \hat{\Pi}^r(\hat{K}), \hat{\Sigma})$ *where* $\hat{K}$ *is a standard simplex in* $\mathbb{R}^d$ *defined by* (7.18)*,* $\hat{\Pi}^r(\hat{K})$ *is the space of polynomials on* $\hat{K}$ *of degree not greater than* r*, and* $\hat{\Sigma}$ *is a set of unisolvent functionals on* $\hat{\Pi}^r(\hat{K})$.

Mappings G_K: $\hat{K} \to K$ *defined for each* $K \in \mathcal{T}_h$ *by the expression* $G_K(\hat{x}) = J_K\hat{x} + b_K$, $\hat{x} \in \hat{K}$*, where* J_K *is a nonsingular matrix and* b_K *is a vector, are called affine mappings. The collection of finite elements* $(K, \Pi^r(K), \Sigma)$ *for* $K \in \mathcal{T}_h$ *is an* affine family of finite elements *if*

(i) $K = G_K(\hat{K})$*;*

(ii) $\Pi^r(K) = \{P\colon\ P = \hat{P} \circ G_K^{-1}, \hat{P} \in \hat{\Pi}^r(\hat{K})\}$*;*

(iii) $\Sigma = \{\sigma_i\colon\ \sigma_i(P(x)) = \hat{\sigma}_i\big(P(G_K(\hat{x}))\big), \hat{\sigma}_i \in \hat{\Sigma}, x = G_K(\hat{x})\}$.

LEMMA. 7.9 *Let $\mathcal{T}_h$ be a family of regular triangulations of affine finite elements. For any integer $m \geq 0$ and each $v \in H^m(K)$ we define a function $\hat{v} = v \circ G_K$: $\hat{K} \to \mathbb{R}$. Then $\hat{v} \in H^m(\hat{K})$ and the following estimates hold*

$$|\hat{v}|_{H^m(\hat{K})} \leq C\|J_K\|^m |\det J_K|^{-\frac{1}{2}} |v|_{H^m(K)}, \tag{7.19}$$

$$|v|_{H^m(K)} \leq C\|J_K^{-1}\|^m |\det J_K|^{\frac{1}{2}} |\hat{v}|_{H^m(\hat{K})}, \tag{7.20}$$

with the constant C depending only on m.

The operator norm $\|\cdot\|$ is implied by the Euclidean norm in $\mathbb{R}^d$ and we have the following estimates

$$\|J_K\| \leq \frac{h_K}{\rho(\hat{K})}, \quad \|J_K^{-1}\| \leq \frac{h_{\hat{K}}}{\rho(K)}. \tag{7.21}$$

Proof. Let us recall that

$$|\hat{v}|_{H^m(\hat{K})} = \Big(\sum_{|\alpha|=m} \int_{\hat{K}} |D^\alpha \hat{v}|^2 d\hat{x} \Big)^{\frac{1}{2}}.$$

By the definition of $\hat{v}$ and the chain rule we have for $|\alpha| = m$

$$\begin{aligned} \|D^\alpha \hat{v}\|_{L^2(\hat{K})} &\leq C\|J_K\|^m \sum_{|\beta|=m} \|(D^\beta v) \circ G_K\|_{L^2\hat{K})} \\ &\leq C\|J_K\|^m |\det J_K|^{-\frac{1}{2}} \sum_{|\beta|=m} \|D^\beta v\|_{L^2(K)}. \end{aligned}$$

Summing the last inequality over all multi-indices α with $|\alpha| = m$ we get (7.19). The proof of (7.20) is analogous.

To prove estimates (7.21) let us recall the definition of the operator norm

$$\|J_K\| = \frac{1}{\rho(\hat{K})} \sup_{|\xi|=\rho(\hat{K})} |J_K \xi|.$$

For each ξ such that $|\xi| = \rho(\hat{K})$, we can find two points $\hat{x}, \hat{y} \in \hat{K}$ with $\hat{x} - \hat{y} = \xi$. Since $J_K \xi = G_K(\hat{x}) - G_K(\hat{y})$ we have the estimate $|J_K \xi| \leq h_K$ which gives the first of estimates (7.21). The proof of the second is similar. ∎

Similarly to the one-dimensional case, we define the spaces

$$X_h^r = \{v_h \in C(\bar{U}) \colon v_h\big|_K \in \Pi^r, \quad \forall K \in \mathcal{T}_h\},$$

and the finite element space $V_h = X_h^r \cap H_0^1(U)$. For X_h^r we define the interpolation operator

$$I_h^r \colon C(\bar{U}) \to X_h^r,$$

which for a given triangulation $\mathcal{T}_h$ of U is defined on each finite element $K \in \mathcal{T}_h$ as the local interpolation operator associating to a continuous function v the polynomial $I_h^r v \in \Pi^r(K)$ such that

$$I_h^r v(x) = \sum_{i=1}^{M_r} v(a_i)\phi_i(x),$$

where $\{\phi_i\}$ is a basis of $\Pi^r(K)$ and $\{a_i\}$ are nodes in K.

To prove the estimate of the interpolation error, we need the following improved version of the Bramble-Hilbert lemma [15].

LEMMA. 7.10 *Let U be a bounded, convex domain in $\mathbb{R}^d$. If $v \in H^m(U)$, $m \geq 1$, then for $0 \leq j < m$*

$$\inf_{Q \in \Pi^{m-1}(U)} |v - Q|_{H^j(U)} \leq C|v|_{H^m(U)},$$

where the constant $C = C(d, m, j, U)$.

We begin with a local estimate of the interpolation error.

THEOREM. 7.11 *Let $\mathcal{T}_h$ be a family of regular triangulations of affine finite elements in $\mathbb{R}^d$. Let $2(r+1) > d$ and $0 \leq m \leq r+1$. Then there exists a constant $C = C(r, m, \hat{K})$ such that*

$$|v - I_h^r v|_{H^m(K)} \leq C \frac{h_K^{r+1}}{\rho^m(K)} |v|_{H^{r+1}(K)}, \quad \forall v \in H^{r+1}(K).$$

Proof. Since K is bounded and convex, we have the Sobolev embedding (Theorem 5.15) for $2(r+1-s) > d$

$$H^{r+1}(K) = H^{(r+1-s)+s}(K) \subsetneq C^s(K).$$

The interpolation operator I_h^r is then well defined in $H^{r+1}(K)$.

Due to the above Sobolev embedding we have the interpolation operator $\hat{I}_h^r \colon C^s(\hat{K}) \to H^{r+1}(\hat{K})$ defined by the expression $I_h^r v \circ G_K = \hat{I}_h^r \hat{v}$. An immediate consequence of this definition is the following equality

$$|(v - I_h^r v) \circ G_K|_{H^m(\hat{K})} = |\hat{v} - \hat{I}_h^r \hat{v}|_{H^m(\hat{K})}. \tag{7.22}$$

Since $\hat{I}_h^r \hat{v}(\hat{x}) = \sum_{i=1}^{M_r} \hat{v}(\hat{a}_i)\hat{\phi}_i(\hat{x})$ the boundedness of this operator follows from the estimate

$$\|\hat{I}_h^r \hat{v}\|_{H^{r+1}(\hat{K})} \leq \sum_{i=1}^{M_r} |\hat{v}(\hat{a}_i)| \|\hat{\phi}_i\|_{H^{r+1}(\hat{K})} \leq C \sup_{\hat{x}} |\hat{v}(\hat{x})| = C\|\hat{v}\|_{C(\hat{K})}.$$

We begin the proof of the theorem showing that for all $0 \leq j \leq r+1$ the following estimate holds

$$|\hat{v} - \hat{I}_h^r \hat{v}|_{H^j(\hat{K})} \leq C|\hat{v}|_{H^{r+1}(\hat{K})}. \tag{7.23}$$

Taking a polynomial $\hat{Q} \in \Pi^r(\hat{K})$ we obtain by the boundedness of $\hat{I}_h^r$ and the Sobolev embedding

$$\begin{aligned}
\|\hat{v} - \hat{I}_h^r \hat{v}\|_{H^{r+1}(\hat{K})} &= \|\hat{v} - \hat{Q} - \hat{I}_h^r(\hat{v} - \hat{Q})\|_{H^{r+1}(\hat{K})} \\
&\leq \|\hat{v} - \hat{Q}\|_{H^{r+1}(\hat{K})} + \|\hat{I}_h^r(\hat{v} - \hat{Q})\|_{H^{r+1}(\hat{K})} \\
&\leq \|\hat{v} - \hat{Q}\|_{H^{r+1}(\hat{K})} + C\|\hat{v} - \hat{Q}\|_{C(\hat{K})} \\
&\leq C\|\hat{v} - \hat{Q}\|_{H^{r+1}(\hat{K})}.
\end{aligned}$$

By Lemma 7.10 we have for $j \leq r$

$$|\hat{v} - \hat{Q}|_{H^j(\hat{K})} \leq C|\hat{v}|_{H^{r+1}(\hat{K})}.$$

For $j = r+1$ we get

$$|\hat{v} - \hat{Q}|_{H^{r+1}(\hat{K})} = |\hat{v}|_{H^{r+1}(\hat{K})}$$

as $\hat{Q}$ is a polynomial of degree r.

Since $\hat{Q}$ is arbitrary, then we obtain

$$\|\hat{v} - \hat{I}_h^r \hat{v}\|_{H^{r+1}(\hat{K})} \leq C|\hat{v}|_{H^{r+1}(\hat{K})}.$$

Estimate (7.23) follows immediately from the above inequality.

By Lemma 7.9, equality (7.22), and estimate (7.23) we have

$$\begin{aligned}
|v - I_h^r v|_{H^m(K)} &\leq C\|J_K^{-1}\|^m |\det J_K|^{\frac{1}{2}} |\hat{v} - \hat{I}_h^r \hat{v}|_{H^m(\hat{K})} \\
&\leq C \frac{h_{\hat{K}}^m}{\rho^m(K)} |\det J_K|^{\frac{1}{2}} |\hat{v} - \hat{I}_h^r \hat{v}|_{H^m(\hat{K})} \\
&\leq C \frac{h_{\hat{K}}^m}{\rho^m(K)} |\det J_K|^{\frac{1}{2}} |\hat{v}|_{H^{r+1}(\hat{K})} \\
&\leq C \frac{1}{\rho^m(K)} |\det J_K|^{\frac{1}{2}} |\hat{v}|_{H^{r+1}(\hat{K})},
\end{aligned}$$

where the constant C is a function of $(r, m, \hat{K})$.

Applying (7.19) and (7.21) to the above inequality we obtain

$$|v - I_h^r v|_{H^m(K)} \le C \frac{1}{\rho^m(K)} |\det J_K|^{\frac{1}{2}} |\hat{v}|_{H^{r+1}(\hat{K})}$$
$$\le C \frac{1}{\rho^m(K)} \|J_K\|^{r+1} |v|_{H^{r+1}(K)} \le C \frac{h_K^{r+1}}{\rho^m(K)} |v|_{H^{r+1}(K)},$$

which completes the proof. ∎

Finally, we have the following global estimate.

THEOREM. 7.12 *Let $\{\mathcal{T}_h\}_{h>0}$ be a family of regular triangulations of affine finite elements in a convex domain $U \subset \mathbb{R}^d$ and $v \in H^{r+1}(U)$, for $2(r+1) > d$. Then for $m = 0, 1$ the following estimates hold*

$$|v - I_h^r v|_{H^m(U)} \le C \left(\sum_{K \in \mathcal{T}_h} h_K^{2(r+1-m)} |v|^2_{H^{r+1}(K)} \right)^{\frac{1}{2}}, \tag{7.24}$$

$$|v - I_h^r v|_{H^m(U)} \le C\, h^{r+1-m} |v|_{H^{r+1}(U)}, \tag{7.25}$$

where $C = C(r, m, d)$. The restriction $m = 0, 1$ is the result of the requirement $I_h^r v \in H^m(\Omega)$ which holds only for $m \le 1$.

Proof. By Theorem 7.11 we have

$$|v - I_h^r v|^2_{H^m(U)} = \sum_{K \in \mathcal{T}_h} |v - I_h^r v|^2_{H^m(K)}$$
$$\le C \sum_{K \in \mathcal{T}_h} \frac{h_K^{2(r+1)}}{\rho^{2m}(K)} |v|^2_{H^{r+1}(K)} = C \sum_{K \in \mathcal{T}_h} \frac{h_K^{2m}}{\rho^{2m}(K)} h_K^{2(r+1-m)} |v|^2_{H^{r+1}(K)}$$
$$\le C \sum_{K \in \mathcal{T}_h} h_K^{2(r+1-m)} |v|^2_{H^{r+1}(K)},$$

which proves (7.24) with $C = C(r, m, d)$. Estimate (7.25) follows from the fact that simplices K are not overlapping, then for every $k \ge 0$

$$|v|_{H^k(U)} = \left(\sum_{K \in \mathcal{T}_h} |v|^2_{H^k(K)} \right)^{\frac{1}{2}}.$$

∎

Let $V_h = X_h^r \cap H_0^1(U)$. We can introduce the following finite element problem for the second-order elliptic equation (7.1) with the Dirichlet boundary conditions

$$\text{find } u_h \in V_h\colon\ B[u_h, v_h] = (f, v_h), \quad \forall v_h \in V_h. \tag{7.26}$$

The existence of a unique u_h which solves (7.26) follows from the equivalence of (7.26) with (7.6) and the unique solvability of the linear system (7.7). Once the interpolation error is estimated, we can estimate the approximation errors similarly to the one-dimensional case. Then we have the following theorems.

THEOREM. 7.13 *Let $u \in H_0^1(U)$ be a solution of* (7.3)*, $U \subset \mathbb{R}^d$, and $u_h \in V_h$ its approximate finite element solution, i.e., a solution of* (7.26)*, obtained with affine finite elements of degree r for a family of regular triangulations $\mathcal{T}_h$. If $u \in H^{r+1}(U)$, $2(r+1) > d$, then the following* a priori *error estimates hold*

$$\begin{aligned} \|u - u_h\|_{H_0^1(U)} &\le \frac{\alpha}{\beta} C \Bigg(\sum_{K \in \mathcal{T}_h} h_K^{2r} |u|^2_{H^{r+1}(K)} \Bigg)^{\frac{1}{2}}, \\ \|u - u_h\|_{H_0^1(U)} &\le \frac{\alpha}{\beta} C\, h^r |u|_{H^{r+1}(U)}, \end{aligned} \tag{7.27}$$

where the constant C is independent of h and u.

THEOREM. 7.14 *Let $u \in H_0^1(U)$ be a solution of* (7.3)*, $U \subset \mathbb{R}^d$, and $u_h \in V_h$ its approximate solution, i.e., a solution of* (7.26)*, obtained with affine finite elements of degree r for a family of regular triangulations $\mathcal{T}_h$. If $u \in H^{m+1}(U) \cap C(\bar{U})$, for $m > 0$, then the following* a priori *error estimate holds*

$$\|u - u_h\|_{L^2(U)} \le C\, h^{s+1} |u|_{H^{s+1}(U)}, \quad s = \min(m, r) \tag{7.28}$$

where the constant C is independent of h and u.

We conclude this section with some remarks on error control. Due to the estimates of Theorem 7.13, we can reduce the approximation error by decreasing the size of the finite element mesh or increasing r. The increase of r is not very common as polynomials of higher-order lead to unstable numerical schemes. Hence, to reduce the approximation error, it is advised to refine the grid. But decreasing h for the whole grid is not the most efficient strategy. It is more convenient to rely on the first of estimates (7.27) and refine the grid only for those elements K on which the contribution to the global error is large. This strategy is called *grid adaptivity*, and its implementation requires the grid refinement for elements K on which the seminorm $|u|_{H^{r+1}(K)}$ is large. This is called *a priori adaptivity*

as we use a priori estimates. The strategy seems to be of little use as the exact solution u is not known. We can, however, replace u in the a priori estimate by a well-chosen approximation. That strategy is, however, not that simple as even for a two-dimensional problem, we need the seminorm in H^2. Assume that we use linear elements. Then u_h is continuous and even piecewise differentiable. But to compute the approximate second derivative of u_h is not a straightforward operation since we need a special reconstruction technique for the first-order derivatives. We are not going into details of this procedure, but it is clear that using grid adaptivity we need sophisticated software for automatic grid generation.

In use is also *a posteriori adaptivity* since it is possible to obtain an error estimate in terms of the approximate solution u_h, so-called *a posteriori estimates*. That approach is not much easier for a computer implementation since we replace the reconstruction technique with the numerically demanding *a posteriori* estimates.

7.2 Finite elements for parabolic equations

We consider now the linear parabolic initial-boundary value problem

$$\begin{aligned} &\tfrac{\partial}{\partial t}u + \mathcal{A}^t u = f, \quad \text{in } (0,T]\times U,\\ &u = 0, \quad \text{on } [0,T]\times \partial U,\\ &u = g, \quad \text{on } \{t=0\}\times U, \end{aligned} \tag{7.29}$$

where $U \subset \mathbb{R}^d$ and $\mathcal{A}^t$ is the uniformly parabolic operator in divergence form defined in Section 5.3. To simplify the presentation we assume that the coefficients of $\mathcal{A}^t$ are time independent. Similarly to the elliptic case we assume that the bilinear form generated by operator $\mathcal{A}^t$ is coercive and is given by equation (7.2). Then we drop the superscript t from $\mathcal{A}^t$ and $B^t[u,v]$ since these operators are independent of time. By a weak solution of problem (7.29) we mean a function $u(t) = u(t,\cdot)$, $u(t) \in H_0^1(U)$ for $t \in [0,T]$, which fulfills the following initial value problem

$$\begin{aligned} &\langle \tfrac{d}{dt}u(t), v\rangle + B[u(t),v] = (f(t),v), \quad \forall v \in H_0^1(U),\\ &u(0) = g. \end{aligned} \tag{7.30}$$

To find a numerical solution we approximate (7.30) by finite elements in space variables. Let V_h be an M_r-dimensional space of finite elements and $\{\phi_i\}_{i=1}^{M_r}$ a basis in V_h. First, we consider the so-called *spatially semi-discrete problem*: find a function $u_h(t) = u_h(t,\cdot)$ such that $u_h(t) \in V_h$ for all $t \in [0,T]$ and u_h is a solution of the following initial value problem

$$\begin{aligned} &\langle \tfrac{d}{dt}u_h(t), v\rangle + B[u_h(t),v] = (f(t),v), \quad \forall v \in V_h,\\ &u_h(0) = g_h, \end{aligned} \tag{7.31}$$

where $g_h \in V_h$ is an approximation of g.

The function u_h can be expanded in the basis of V_h

$$u_h(t,x) = \sum_{i=1}^{M_r} w_i(t)\phi_i(x).$$

Inserting this series into (7.31), we obtain

$$\sum_{i=1}^{M_r} \tfrac{d}{dt} w_i(t)(\phi_i,\phi_j) + \sum_{i=1}^{M_r} w_i(t) B[\phi_i,\phi_j] = (f(t),\phi_j), \quad j = 1,\dots,M_r.$$

Denoting by $W(t) = \big(w_1(t),\dots,w_{M_r}(t)\big)$ the vector of unknowns, by E_h the mass matrix with the elements

$$\big(E_h\big)_{ij} = (\phi_i,\phi_j), \quad i,j = 1,\dots,M_r,$$

by A_h the stiffness matrix with the elements

$$\big(A_h\big)_{ij} = B[\phi_i,\phi_j], \quad i,j = 1,\dots,M_r,$$

and by $F_h(t) = \big(f_1(t),\dots,f_{M_r}(t)\big)$ the load vector with the components

$$f_i(t) = (f(t),\phi_i), \quad i = 1,\dots,M_r,$$

we reduce the Galerkin approximation to the following system of ordinary differential equations

$$E_h \tfrac{dW(t)}{dt} + A_h W(t) = F_h(t). \tag{7.32}$$

Let us observe that E_h is a positive definite matrix since $\{\phi_i\}$ is a basis of V_h. A_h is also a positive definite matrix by the coerciveness of $B[u,v]$. Hence the system (7.32) possesses a unique solution and also u_h which solves (7.31) is uniquely defined.

To obtain a fully discretized problem we can apply various discretizations. The most popular one is to approximate the time derivative by finite differences and apply two-time-level schemes. Hence, we divide the time interval $[0,T]$ into N subintervals, set $\delta t = \frac{T}{N}$ and obtain the time grid $t_n = n\cdot\delta t$, $n = 0,1,\dots,N$. Denote $W^n = W(t_n)$ and $F^n = F_h(t_n)$. We will consider the approximation of system (7.32) by the θ-scheme

$$E_h \frac{W^{n+1}-W^n}{\delta t} + A_h\big(\theta W^{n+1} + (1-\theta)W^n\big) = \big(\theta F^{n+1} + (1-\theta)F^n\big). \tag{7.33}$$

Then we define the fully discrete Galerkin approximation

$$u_h^n(x) = \sum_{i=1}^{M_r} w_i^n \phi_i(x), \tag{7.34}$$

where $w_i^n = w_i(t_n)$ according to the previously introduced notation.

Our goal is to prove that u_h^n converge to $u(t,x)$ a solution of the continuous problem (7.29). We begin with the proof of stability for the θ-schemes. An important step in the proof is based on the following *inverse estimate*.

LEMMA. 7.15 *Let $\mathcal{T}_h$ be a family of regular triangulations of $U \subset \mathbb{R}^d$. Then there exists a constant C_{inv} such that for all $v \in V_h$*

$$\|Dv\|_{L^2(U)} \leq C_{inv}\, h^{-1} \|v\|_{L^2(U)},$$

where Dv denotes the gradient of v.

Proof. It is sufficient to prove the above estimate on a single finite element $K \in \mathcal{T}_h$

$$\int_K |Dv|^2 dx \leq C\, h^{-2} \int_K |v|^2 dx.$$

Let $\hat{K}$ be a standard simplex given by formula (7.18) and $G\colon\ \hat{K} \to K$ an affine map. We define $\hat{v}(\hat{x}) = v(G(\hat{x}))$ for $\hat{x} \in \hat{K}$. Denoting $J = \left(\dfrac{\partial x}{\partial \hat{x}}\right)$ the Jacobian matrix of G we get $D\hat{v} = JDv$. By the regularity of triangulation $\|J\| \sim h$, $\|J^{-1}\| \sim h^{-1}$ and we have

$$|D\hat{v}| \leq Ch|Dv|, \quad |Dv| \leq Ch^{-1}|D\hat{v}|.$$

By the change of variables $dx = |\det J| d\hat{x} \simeq h^d d\hat{x}$ we obtain

$$\int_K |Dv|^2 dx \leq C\, h^{d-2} \int_{\hat{K}} |D\hat{v}|^2 d\hat{x}.$$

Let us observe now that, since V_h is a finite-dimensional space, all norms in V_h are equivalent. In particular, the H^k-norm is equivalent to the L^2-norm. Then

$$\int_U \sum_{|\alpha| \leq k} |D^\alpha v|^2 dx \leq C \int_U |v|^2 dx, \quad \forall v \in V_h.$$

Applying that inequality to the H^1 norm we obtain

$$\begin{aligned}\int_K |Dv|^2 dx &\leq C\, h^{d-2} \int_{\hat{K}} |D\hat{v}|^2 d\hat{x} \\ &\leq C\, h^{d-2} \int_{\hat{K}} \sum_{|\alpha| \leq 1} |D^\alpha \hat{v}|^2 d\hat{x} \leq C\, h^{d-2} \int_{\hat{K}} |\hat{v}|^2 d\hat{x} \leq C\, h^{-2} \int_K |v|^2 dx.\end{aligned}$$

■

THEOREM. 7.16 *Assume that the bilinear form $B[u,v]$ fulfills conditions* (7.4) *and* (7.5) *and $\|f(t)\|_{L^2(U)}$ is bounded on $[0,T]$. For $\theta < \frac{1}{2}$ we assume additionally that the following restriction on the time step holds*

$$\delta t(1 + C_{inv}h^{-1})^2 < \frac{2\beta}{(1-2\theta)\alpha^2}, \tag{7.35}$$

where C_{inv} is the constant from the estimate in Lemma 7.15 and α, β are the constants from (7.4) *and* (7.5).

Then u_h^n which solve the weak formulation of the fully discrete approximation

$$\begin{aligned} \frac{1}{\delta t}\big(u_h^{n+1} - u_h^n, v_h\big) &= -B\big[\theta u_h^{n+1} + (1-\theta)u_h^n, v_h\big] \\ &\quad + \big(\theta f(t_{n+1}) + (1-\theta)f(t_n), v_h\big), \quad \forall v_h \in V_h, \\ u_h^0 &= I_h^1 g, \end{aligned} \tag{7.36}$$

have the estimate

$$\|u_h^n\|_{L^2(U)} \le C\bigg(\|u_h^0\|_{L^2(U)} + \delta t \sum_{i=0}^{n} \|f(t_i)\|_{L^2(U)}\bigg),$$

where the constant C depends on θ but is independent of h, δt and N, and is a non-decreasing function of α, β^{-1} and T.

Proof. Taking $v_h = \theta u_h^{n+1} + (1-\theta)u_h^n$ in (7.36) and applying the elementary identity

$$(u_h^{n+1} - u_h^n, u_h^n) + \tfrac{1}{2}(u_h^{n+1} - u_h^n, u_h^{n+1} - u_h^n) = \tfrac{1}{2}\|u_h^{n+1}\|_{L^2(U)} - \tfrac{1}{2}\|u_h^n\|_{L^2(U)},$$

we obtain

$$\begin{aligned} \tfrac{1}{2}\|u_h^{n+1}\|^2_{L^2(U)} - \tfrac{1}{2}\|u_h^n\|^2_{L^2(U)} + (\theta - 1/2)\|u_h^{n+1} - u_h^n\|^2_{L^2(U)} \\ + \delta t B\big[\theta u_h^{n+1} + (1-\theta)u_h^n, \theta u_h^{n+1} + (1-\theta)u_h^n\big] \\ \le \delta t\big(\theta f(t_{n+1}) + (1-\theta)f(t_n), \theta u_h^{n+1} + (1-\theta)u_h^n\big). \end{aligned}$$

By the coerciveness of $B[u,v]$ and the Cauchy inequality, we get for $0 < \epsilon \le 1$

$$\begin{aligned} \|u_h^{n+1}\|^2_{L^2(U)} - \|u_h^n\|^2_{L^2(U)} + (2\theta - 1)\|u_h^{n+1} - u_h^n\|^2_{L^2(U)} \\ + 2(1-\epsilon)\beta\delta t\|\theta u_h^{n+1} + (1-\theta)u_h^n\|^2_{H_0^1(U)} \\ \le \frac{\delta t}{2\epsilon\beta}\|\theta f(t_{n+1}) + (1-\theta)f(t_n)\|^2_{H^{-1}(U)}. \end{aligned} \tag{7.37}$$

For $\theta \geq \frac{1}{2}$, (7.37) gives the estimate of $\|u_h^{n+1}\|^2_{L^2(U)} - \|u_h^n\|^2_{L^2(U)}$ which is sufficient to complete the proof. But for $\theta < \frac{1}{2}$, we need a better estimate. To this end, we insert into (7.36) $v_h = u_h^{n+1} - u_h^n$ to obtain

$$\begin{aligned}\|u_h^{n+1} - u_h^n\|^2_{L^2(U)} &= -\delta t B\big[\theta u_h^{n+1} + (1-\theta)u_h^n, u_h^{n+1} - u_h^n\big] \\ &\quad + \delta t\big(\theta f(t_{n+1}) + (1-\theta)f(t_n), u_h^{n+1} - u_h^n\big) \\ &\leq \alpha\delta t\|\theta u_h^{n+1} + (1-\theta)u_h^n\|_{H_0^1(U)}\|u_h^{n+1} - u_h^n\|_{H_0^1(U)} \\ &\quad + \delta t\|\theta f(t_{n+1}) + (1-\theta)f(t_n)\|_{H^{-1}(U)}\|u_h^{n+1} - u_h^n\|_{H_0^1(U)}.\end{aligned}$$

By Lemma 7.15 we have

$$\begin{aligned}\|u_h^{n+1} - u_h^n\|_{L^2(U)} \leq &\delta t(1 + C_{inv}h^{-1})\Big(\alpha\|\theta u_h^{n+1} + (1-\theta)u_h^n\|_{H_0^1(U)} \\ &+ \|\theta f(t_{n+1}) + (1-\theta)f(t_n)\|_{H^{-1}(U)}\Big).\end{aligned} \tag{7.38}$$

Inserting this estimate into (7.37) and defining for $\eta > 0$ the constant

$$\kappa = 2(1-\epsilon)\beta - (1-2\theta)\alpha^2(1+\eta)\delta t(1 + C_{inv}h^{-1})^2$$

we obtain

$$\begin{aligned}\|u_h^{n+1}\|^2_{L^2(U)} - \|u_h^n\|^2_{L^2(U)} + \delta t\kappa\|\theta u_h^{n+1} + (1-\theta)u_h^n\|^2_{H_0^1(U)} \\ \leq C\,\delta t(1 + \delta t\,h^{-2})\|\theta f(t_{n+1}) + (1-\theta)f(t_n)\|^2_{H^{-1}(U)}.\end{aligned} \tag{7.39}$$

Let us observe that for sufficiently small ϵ, η and $\theta < \frac{1}{2}$, we have $0 < \kappa < +\infty$ due to (7.35). This gives the estimate of $\|u_h^{n+1}\|^2_{L^2(U)} - \|u_h^n\|^2_{L^2(U)}$.

Let m be fixed, $1 \leq m \leq N$. Summing up the estimates for $\theta < \frac{1}{2}$ from $n = 0$ to $n = m-1$ and similarly the estimates for $\theta \geq \frac{1}{2}$ we obtain

$$\|u_h^m\|^2_{L^2(U)} \leq \|u_h^0\|^2_{L^2(U)} + C\,\delta t\sum_{n=0}^{m-1}\|\theta f(t_{n+1}) + (1-\theta)f(t_n)\|^2_{H^{-1}(U)}.$$

Since $\|\cdot\|_{H^{-1}} \leq \|\cdot\|_{L^2}$ the assertion of the theorem follows. ∎

The proof of convergence of the finite element approximation will go in two steps. First, we prove an error estimate for the semi-discrete problem (7.31). The convergence will be carried on under the simplified assumption that the bilinear form $B[u, v]$ is symmetric. That corresponds to the assumption that the operator $\mathcal{A}$ in problem (7.29) is self-adjoint.

DEFINITION. 7.17 *Let the bilinear form $B[u,v]$ implied by the operator $\mathcal{A}$ from* (7.29) *be symmetric and fulfill estimates* (7.4) *and* (7.5). *In $H_0^1(U)$ we define the* Ritz projection R_h

$$R_h\colon\ H_0^1(U) \to V_h$$

by the equality

$$B[R_h u, v_h] = B[u, v_h], \quad \forall u \in H_0^1(U),\ \forall v_h \in V_h.$$

The properties of the Ritz projection most relevant for the proofs of subsequent theorems are summarized in the following lemma.

LEMMA. 7.18

1. *For each $u \in H_0^1(U)$ and $v_h \in V_h$ the following orthogonality condition holds*
$$B[R_h u - u, v_h] = 0.$$

2. *For all $u \in H_0^1(U)$ we have*
$$\|R_h u - u\|_{\mathcal{H}} = \min_{v_h \in V_h} \|v_h - u\|_{\mathcal{H}}$$
where $\|v\|_{\mathcal{H}} = \big(B[v,v]\big)^{\frac{1}{2}}$ is a norm in $H_0^1(U)$ equivalent to the usual norm of this space due to the symmetry of the bilinear form $B[u,v]$, and estimates (7.4) *and* (7.5).

3. *If u is a weak solution of* (7.3) *then the Ritz projection $R_h u = u_h$, where u_h is a finite element solution of* (7.26).

Proof. Condition 1. follows from the definition of the Ritz projection. To prove Condition 2. let us recall that due to (7.4) and (7.5) $\big(B[\cdot,\cdot]\big)^{\frac{1}{2}}$ is a norm in $H_0^1(U)$ equivalent to the norm $\|\cdot\|_{H_0^1(U)}$. Condition 1. implies that R_h is the orthogonal projection of $H_0^1(U)$ on V_h in the norm $\|\cdot\|_{\mathcal{H}}$. Then Condition 2. is a property of orthogonal projections in a Hilbert space.

To prove Condition 3. let us observe that the orthogonality relation for the Ritz projection is in fact the orthogonality relation (7.16) which is valid for the finite element solution u_h. Since $R_h u$ is uniquely defined by the fact that R_h is an orthogonal projection in $H_0^1(U)$ then $R_h u = u_h$. ∎

The following lemma extends the error estimates for the interpolation operator in V_h to the Ritz projection.

LEMMA. 7.19 *Let U be an open, convex, bounded domain in $\mathbb{R}^d$. Then for $m = 1, 2$ we have the estimate*

$$\|R_h u - u\|_{L^2(U)} + h|R_h u - u|_{H_0^1(U)} \le C\, h^m \|u\|_{H^m(U)}, \quad \forall u \in H^m(U) \cap H_0^1(U).$$

Proof. Let us recall that due to the assumptions on $B[u, v]$ and the Poincaré inequality for a bounded U we have the inequalities

$$\begin{aligned} &C_1\|v\|_{H_0^1(U)} \le \|v\|_{\mathcal{H}} \le C_2\|v\|_{H_0^1(U)}, \\ &C_1|v|_{H_0^1(U)} \le \|v\|_{H_0^1(U)} \le C_2|v|_{H_0^1(U)}. \end{aligned} \tag{7.40}$$

The H_0^1 estimate of the lemma follows from the fact that R_h is an orthogonal projection in $\mathcal{H}$. Then $\|R_h u\|_{\mathcal{H}} \le \|u\|_{\mathcal{H}}$ and by inequalities (7.40) $|R_h u|_{H_0^1(U)} \le C|u|_{H_0^1(U)}$. Hence

$$|R_h u - u|_{H_0^1(U)} \le C|u|_{H_0^1(U)} \le C\|u\|_{H_0^1(U)}. \tag{7.41}$$

Since the Ritz projection is an orthogonal projection $\mathcal{H} \to V_h$ we have

$$\|R_h u - u\|_{\mathcal{H}} = \min_{v \in V_h} \|v - u\|_{\mathcal{H}},$$

which gives

$$\|R_h u - u\|_{\mathcal{H}} \le \|I_h^1 u - u\|_{\mathcal{H}}.$$

Then by (7.40)

$$|R_h u - u|_{H_0^1(U)} \le C\|R_h u - u\|_{\mathcal{H}} \le C\|I_h^1 u - u\|_{H_0^1(U)}.$$

Hence by (7.40) and the estimates of Theorem 7.12 we get

$$|R_h u - u|_{H_0^1(U)} \le C\,|I_h^1 u - u|_{H_0^1(U)} \le C\,h|u|_{H^2(U)} \le C\,h\|u\|_{H^2(U)}. \tag{7.42}$$

To obtain the L^2 estimate of $(R_h u - u)$ we consider the following weak dual problem with the right hand side $e = R_h u - u$

$$\text{find } g \in H_0^1(U):\ B[v, g] = (v, e), \quad \forall v \in H_0^1(U).$$

Due to the assumptions on B, this problem possesses a unique weak solution and the following regularity estimate holds (cf. Theorem 5.22)

$$\|g\|_{H^2(U)} \le C\|e\|_{L^2(U)}.$$

Taking $v = e$, using the orthogonality of the Ritz projection (Condition 1. of Lemma 7.18), estimates (7.4) and (7.40), the estimates of Theorem 7.12 and the above regularity estimate we get

$$\begin{aligned}\|e\|^2_{L^2(U)} = B[e,g] = B[e, g - I^1_h g] &\le C\,\|e\|_{H^1_0(U)}\|g - I^1_h g\|_{H^1_0(U)} \\ &\le C\,h\|e\|_{H^1_0(U)}\|g\|_{H^2(U)} \le C\,h\|e\|_{H^1_0(U)}\|e\|_{L^2(U)}.\end{aligned}$$

Hence $\|e\|_{L^2(U)} \le C\,h\|e\|_{H^1_0(U)}$ and since by (7.40) and (7.42) we have

$$\|R_h u - u\|_{H^1_0(U)} \le C\,|R_h u - u|_{H^1_0(U)} \le C\,h\|u\|_{H^2(U)},$$

that completes the proof for $m = 2$.

Since for $m = 2$, we have proven the estimate

$$\|e\|_{L^2(U)} \le C\,h\|e\|_{H^1_0(U)} \le C\,h|e|_{H^1_0(U)}$$

then by (7.41) we get

$$\|R_h u - u\|_{L^2(U)} \le C\,h|R_h u - u|_{H^1_0(U)} \le C\,h|u|_{H^1_0(U)} \le C\,h\|u\|_{H^1_0(U)}.$$

■

Remark. 7.1 *The above estimates can be immediately extended to finite elements of order $r > 2$. It is sufficient to apply the estimates of Theorem 7.12 with an appropriate r. We then get, for $2 \le s \le r$*

$$\|R_h u - u\|_{L^2(U)} + h|R_h u - u|_{H^1_0(U)} \le C\,h^s\|u\|_{H^s(U)}, \quad \forall u \in H^s(U) \cap H^1_0(U).$$

This estimates are applicable to solutions of elliptic or parabolic equations in domains with smooth boundaries since only for such domains we can expect that $u \in H^r(U)$ for $r > 2$ (cf. Theorem 5.22 and Remark 5.3). In polyhedral domains, we cannot expect higher regularity of solutions because singularities can develop in the corners of the domain. Hence there is no justification for using finite elements of order higher than 2 in polyhedral domains. For convex domains with smooth boundaries, the regularity of solutions is not a problem but to obtain a high order approximation error a special consideration of the boundary layer $U \setminus U_h$ is needed.

THEOREM. 7.20 *Let $u \in L^2(0,T;H^4(U))$ be a solution of* (7.30) *with $g \in H^3(U)$ and a sufficiently smooth f, and $u_h(t)$ a solution of* (7.31)*. Then for $t \ge 0$*

$$\|u_h(t) - u(t)\|_{L^2(U)} \le \|g_h - g\|_{L^2(U)} + C\,h^2\left(\|g\|_{H^2(U)} + \int_0^t \|\tfrac{du}{ds}\|_{H^2(U)} ds\right).$$

Proof. Using the Ritz projection, we can write

$$u_h - u = (u_h - R_h u) + (R_h u - u).$$

The second term can be estimated using Lemma 7.19

$$\begin{aligned}\|R_h u(t) - u(t)\|_{L^2(U)} &\le C\,h^2 \left\| g + \int_0^t \tfrac{du(s)}{ds} ds \right\|_{H^2(U)} \\ &\le C\,h^2 \Big(\|g\|_{H^2(U)} + \int_0^t \|\tfrac{du}{ds}\|_{H^2(U)} ds \Big).\end{aligned}$$

Let $D(t) = u_h(t) - R_h u(t)$. Then using the definition of the Ritz projection and the fact that R_h commutes with time differentiation we obtain

$$\begin{aligned}\big(\tfrac{dD}{dt}(t), v_h\big) + B[D(t), v_h] &= \big(\tfrac{d}{dt} u_h(t), v_h\big) + B[u_h(t), v_h] - \big(R_h \tfrac{du}{dt}(t), v_h\big) \\ -B[R_h u(t), v_h] &= (f(t), v_h) - \big(R_h \tfrac{du}{dt}(t), v_h\big) - B[R_h u(t), v_h] \\ &= (f(t), v_h) - \big(R_h \tfrac{du}{dt}(t), v_h\big) - B[u(t), v_h] \\ &= \big(\tfrac{du}{dt}(t) - R_h \tfrac{du}{dt}(t), v_h\big).\end{aligned}$$

Taking $v_h = D(t)$ and using the coerciveness of B we obtain

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|D(t)\|^2_{L^2(U)} &\le \big(\tfrac{du}{dt}(t) - R_h \tfrac{du}{dt}(t), D(t)\big) \\ &\le \|D(t)\|_{L^2(U)} \|\tfrac{du}{dt}(t) - R_h \tfrac{du}{dt}(t)\|_{L^2(U)},\end{aligned}$$

which gives the estimate

$$\frac{d}{dt}\|D(t)\|_{L^2(U)} \le \|\tfrac{du}{dt}(t) - R_h \tfrac{du}{dt}(t)\|_{L^2(U)}.$$

Integrating the above inequality, we have

$$\|D(t)\|_{L^2(U)} \le \|D(0)\|_{L^2(U)} + \int_0^t \|\tfrac{du}{ds}(s) - R_h \tfrac{du}{ds}(s)\|_{L^2(U)} ds.$$

Using the estimate from Lemma 7.19 we get

$$\|\tfrac{du}{dt}(t) - R_h \tfrac{du}{dt}(t)\|_{L^2(U)} \le C\,h^2 \|\tfrac{du}{dt}(t)\|_{H^2(U)}$$

and

$$\begin{aligned}\|D(0)\|_{L^2(U)} = \|g_h - R_h g\|_{L^2(U)} &\le \|g_h - g\|_{L^2(U)} + \|R_h g - g\|_{L^2(U)} \\ &\le \|g_h - g\|_{L^2(U)} + c\,h^2 \|g\|_{H^2(U)}.\end{aligned}$$

■

The above result can be extended to finite elements of order $r > 1$. When a weak solution v of (7.3) is a function in $H_0^1(U) \cap H^r(U)$ then due to Remark 7.1 we have the estimate of the Ritz projection

$$\|v - R_h v\|_{L^2(U)} \leq C\, h^r \|v\|_{H^r(U)}. \tag{7.43}$$

In the finite element space V_h that fulfills (7.43), we have a stronger estimate, but that result has limited applicability in polyhedral domains (see Remark 7.1).

THEOREM. 7.21 *Let $u(t)$ be a solution of* (7.30) *and $u_h(t)$ a solution of* (7.31). *Then*

$$\|u_h(t) - u(t)\|_{L^2(U)} \leq C\, h^r \left(\|g_h - g\|_{H^r(U)} + \int_0^t \|\tfrac{du}{ds}\|_{H^r(U)} ds \right),$$

where $g_h \in X_h^r \cap H_0^1(U) \cap H^{r+1}(U)$, $g \in H_0^1(U) \cap H^{r+1}(U)$ and $u(t) \in H_0^1(U) \cap H^{r+1}(U)$, $\frac{du}{dt}(t) \in H_0^1(U) \cap H^r(U)$.

For the fully discrete Galerkin method, the following estimates can be obtained.

THEOREM. 7.22 *Let u_h^n be computed by the implicit Euler scheme (scheme* (7.36) *with $\theta = 1$) with $g_h = I_h^1 g$. Let $u \in L^2(0,T; H^4(U))$ be a solution of* (7.30) *with $g \in H^3(U)$ and a sufficiently smooth f (cf. Remark 5.3). Then*

$$\begin{aligned} \|u_h^n - u(t_n)\|_{L^2(U)} \leq & C\, h^2 \left(\|g\|_{H^2(U)} + \int_0^{t_n} \|\tfrac{du}{ds}\|_{H^2(U)} ds \right) \\ & + C\, \delta t \int_0^{t_n} \|\tfrac{d^2u}{ds^2}\|_{L^2(U)} ds. \end{aligned} \tag{7.44}$$

The interpolant $I_h^1 g$ is well defined only when $U \subset \mathbb{R}^d$, $d \leq 5$ (g is continuous due to Theorem 5.15). For $d > 5$ we have to assume additionally that $g \in H^m(U)$, $2m > d$, and appropriately reformulate the theorem.

Proof. For the implicit Euler scheme u_h^n solves the equation

$$\frac{1}{\delta t}\big(u_h^{n+1} - u_h^n, v_h\big) + B[u_h^{n+1}, v_h] = (f(t_{n+1}), v_h),\ \forall v_h \in V_h. \tag{7.45}$$

Using the Ritz projection we can write

$$\|u_h^n - u(t_n)\|_{L^2(U)} \leq \|u_h^n - R_h u(t_n)\|_{L^2(U)} + \|R_h u(t_n) - u(t_n)\|_{L^2(U)}.$$

From Lemma 7.19 we estimate the second term

$$\|R_h u(t_n) - u(t_n)\|_{L^2(U)} \leq C\, h^2 \bigg(\|g\|_{H^2(U)} + \int_0^{t_n} \|\tfrac{du}{ds}\|_{H^2(U)} ds \bigg).$$

To estimate the first term we define $D_h^n = u_h^n - R_h u(t_n)$. Inserting D_h^n into equation (7.45) we obtain

$$\frac{1}{\delta t}\big(D_h^{n+1} - D_h^n, v_h\big) + B[D_h^{n+1}, v_h] = (e^{n+1}, v_h),\ \forall v_h \in V_h, \tag{7.46}$$

where

$$(e^{n+1}, v_h) = (f(t_{n+1}), v_h) - \frac{1}{\delta t}\big(R_h(u(t_{n+1}) - u(t_n)), v_h\big) - B[u(t_{n+1}), v_h].$$

Equation (7.46) for D_h^n is analogous to equation (7.36) for u_h^n with $\theta = 1$. Hence the stability result of Theorem 7.16 gives

$$\|D_h^n\|_{L^2(U)} \leq C \bigg(\|D_h^0\|_{L^2(U)} + \delta t \sum_{i=1}^{n} \|e^i\|_{L^2(U)} \bigg).$$

For D_h^0 we have the estimate

$$\begin{aligned} \|D_h^0\|_{L^2(U)} = \|g_h - R_h g\|_{L^2(U)} &\leq \|g_h - g\|_{L^2(U)} + \|g - R_h g\|_{L^2(U)} \\ &\leq C\, h^2 \|g\|_{H^2(U)}, \end{aligned}$$

by Lemma 7.19 and Theorem 7.12.

To estimate $\|e^i\|_{L^2(U)}$ let us observe that

$$(f(t_i), v_h) - B[u(t_i), v_h] = (\tfrac{du}{dt}(t_i), v_h).$$

Then

$$\begin{aligned} (e^i, v_h) &= (\tfrac{du}{dt}(t_i), v_h) - \frac{1}{\delta t}\big(R_h(u(t_i) - u(t_{i-1})), v_h\big) \\ &= \Big(\tfrac{du}{dt}(t_i) - \frac{u(t_i) - u(t_{i-1})}{\delta t}, v_h\Big) + \Big((I - R_h)\frac{u(t_i) - u(t_{i-1})}{\delta t}, v_h\Big). \end{aligned}$$

Using the Taylor formula we have

$$\tfrac{du}{dt}(t_i) - \frac{u(t_i) - u(t_{i-1})}{\delta t} = \frac{1}{\delta t}\int_{t_{i-1}}^{t_i} (s - t_{i-1}) \tfrac{d^2u(s)}{ds^2} ds.$$

Since R_h commutes with time differentiation, then

$$(I - R_h)(u(t_i) - u(t_{i-1})) = \int_{t_{i-1}}^{t_i} (I - R_h)\tfrac{du(s)}{ds} ds.$$

Applying Lemma 7.19 we obtain

$$\|(I - R_h)(u(t_i) - u(t_{i-1}))\|_{L^2(U)} \leq C\, h^2 \int_{t_{i-1}}^{t_i} \|\tfrac{du}{ds}\|_{H^2(U)} ds.$$

Then

$$\begin{aligned}\|e^i\|_{L^2(U)} &\leq \frac{1}{\delta t}\left\| \int_{t_{i-1}}^{t_i} (s - t_{i-1})\tfrac{d^2u(s)}{ds^2} ds \right\|_{L^2(U)} + \frac{C\, h^2}{\delta t} \int_{t_{i-1}}^{t_i} \|\tfrac{du}{ds}\|_{H^2(U)} ds \\ &\leq \int_{t_{i-1}}^{t_i} \|\tfrac{d^2u}{ds^2}\|_{L^2(U)} ds + \frac{C\, h^2}{\delta t} \int_{t_{i-1}}^{t_i} \|\tfrac{du}{ds}\|_{H^2(U)} ds.\end{aligned}$$

Together these estimates give

$$\begin{aligned}\|D_h^n\|_{L^2(U)} &\leq C\, h^2 \|g\|_{H^2(U)} + C\, \delta t \sum_{i=1}^{n} \frac{C\, h^2}{\delta t} \int_{t_{i-1}}^{t_i} \|\tfrac{du}{ds}\|_{H^2(U)} ds \\ &\qquad + C\, \delta t \sum_{i=1}^{n} \int_{t_{i-1}}^{t_i} \|\tfrac{d^2u}{ds^2}\|_{L^2(U)} ds \\ &\leq C\left(h^2 \|g\|_{H^2(U)} + h^2 \int_0^{t_n} \|\tfrac{du}{ds}\|_{H^2(U)} ds + \delta t \int_0^{t_n} \|\tfrac{d^2u}{ds^2}\|_{L^2(U)} ds \right).\end{aligned}$$

■

Remark. 7.2 *The theorem holds under the weaker assumption $g \in H^2(U)$. But with this weaker assumption the theorem formulation requires a change of norms in the right hand side of estimate* (7.44) *adequate to the regularity of the time derivatives $\frac{du}{dt}(t) \in H_0^1(U)$ and $\frac{d^2u}{dt^2}(t) \in H^{-1}(U)$ (cf. Theorem 5.22). Besides, the dimension of $U \subset \mathbb{R}^d$ has to be restricted to $d \leq 3$.*

The above theorem remains valid for the θ-schemes with $\theta > \frac{1}{2}$ and the proof is essentially the same. For $\theta < \frac{1}{2}$, we can obtain an analogous theorem imposing an additional assumption controlling the time step similar to (7.35). The result obtained is only first-order accurate in time. To obtain second-order accuracy in time, we have to apply the Crank-Nicolson scheme ($\theta = \frac{1}{2}$). Then we obtain the following theorem.

THEOREM. 7.23 *Let u_h^n be computed by the Crank-Nicolson scheme (scheme (7.36) with $\theta = \frac{1}{2}$) with $g_h = I_h^1 g$. Let $u \in L^2(0,T;H^6(U))$ be a solution of (7.30) with $g \in H^5(U)$ and a sufficiently smooth f. Then*

$$\|u_h^n - u(t_n)\|_{L^2(U)} \leq C\,h^2\bigg(\|g\|_{H^2(U)} + \int_0^{t_n} \|\tfrac{du}{ds}\|_{H^2(U)} ds\bigg) + C\,(\delta t)^2 \int_0^{t_n} \|\tfrac{d^3u}{ds^3}\|_{L^2(U)} ds.$$

The mentioned in Theorem 7.22 restrictions on the dimensionality of U and the smoothness of g apply also in this case.

Proof. The proof goes along the same lines as the proof of Theorem 7.22. The Crank-Nicolson scheme for u_h^n is

$$\frac{1}{\delta t}\big(u_h^{n+1} - u_h^n, v_h\big) + B[\tfrac{1}{2}(u_h^{n+1} + u_h^n), v_h] = (\tfrac{1}{2}(f(t_{n+1}) + f(t_n)), v_h),\ \forall v_h \in V_h, \tag{7.47}$$

Then we split

$$\|u_h^n - u(t_n)\|_{L^2(U)} \leq \|u_h^n - R_h u(t_n)\|_{L^2(U)} + \|R_h u(t_n) - u(t_n)\|_{L^2(U)}$$

and estimate the second term from Lemma 7.19

$$\|R_h u(t_n) - u(t_n)\|_{L^2(U)} \leq C\,h^2\bigg(\|g\|_{H^2(U)} + \int_0^{t_n} \|\tfrac{du}{ds}\|_{H^2(U)} ds\bigg).$$

To estimate the first term we define $D_h^n = u_h^n - R_h u(t_n)$. Inserting D_h^n into equation (7.47) we obtain

$$\frac{1}{\delta t}\big(D_h^{n+1} - D_h^n, v_h\big) + B[\tfrac{1}{2}(D_h^{n+1} + D_h^n), v_h] = (\tfrac{1}{2}(e^{n+1} + e^n), v_h),\ \forall v_h \in V_h,$$

where

$$\begin{aligned}(\tfrac{1}{2}(e^{n+1} + e^n), v_h) =& (\tfrac{1}{2}(f(t_{n+1}) + f(t_n)), v_h) - \frac{1}{\delta t}\big(R_h(u(t_{n+1}) - u(t_n)), v_h\big) \\ & - B[\tfrac{1}{2}(u(t_{n+1}) + u(t_n)), v_h].\end{aligned}$$

Similarly to the proof of Theorem 7.22, we get the estimate of D_h^n

$$\|D_h^n\|_{L^2(U)} \leq C\bigg(\|D_h^0\|_{L^2(U)} + \delta t \sum_{i=1}^{n} \|\tfrac{1}{2}(e^i + e^{i-1})\|_{L^2(U)}\bigg)$$

and D_h^0

$$\begin{aligned}\|D_h^0\|_{L^2(U)} = \|g_h - R_h g\|_{L^2(U)} &\le \|g_h - g\|_{L^2(U)} + \|g - R_h g\|_{L^2(U)} \\ &\le C\, h^2 \|g\|_{H^2(U)}.\end{aligned}$$

To estimate $\|\frac{1}{2}(e^i + e^{i-1})\|_{L^2(U)}$ we use the identity

$$(f(t_i), v_h) - B[u(t_i), v_h] = (\tfrac{du}{dt}(t_i), v_h).$$

Then

$$\begin{aligned}(\tfrac{1}{2}(e^i + e^{i-1}), v_h) &= (\tfrac{1}{2}(\tfrac{du}{dt}(t_i) + \tfrac{du}{dt}(t_{i-1})), v_h) - \frac{1}{\delta t}\big(R_h(u(t_i) - u(t_{i-1})), v_h\big) \\ &= \Big((I - R_h)\frac{u(t_i) - u(t_{i-1})}{\delta t}, v_h\Big) + \Big(\tfrac{du}{dt}(t_{i-1/2}) - \frac{u(t_i) - u(t_{i-1})}{\delta t}, v_h\Big) \\ &\quad + \Big(\frac{1}{2}\big(\tfrac{du}{dt}(t_i) + \tfrac{du}{dt}(t_{i-1})\big) - \tfrac{du}{dt}(t_{i-1/2}), v_h\Big),\end{aligned}$$

where $t_{i-1/2} = \frac{1}{2}(t_i + t_{i-1})$.

For the first term on the right hand side, we get like in Theorem 7.22

$$\|(I - R_h)(u(t_i) - u(t_{i-1}))\|_{L^2(U)} \le C\, h^2 \int_{t_{i-1}}^{t_i} \|\tfrac{du}{ds}\|_{H^2(U)} ds.$$

Integrating by parts two times we have for the second term

$$\begin{aligned}\frac{1}{2}\int_{t_{i-1}}^{t_{i-1/2}} (s - t_{i-1})^2 \tfrac{d^3u(s)}{ds^3} ds + \frac{1}{2}\int_{t_{i-1/2}}^{t_i} (s - t_i)^2 \tfrac{d^3u(s)}{ds^3} ds \\ = (t_i - t_{i-1})\Big(\frac{u(t_i) - u(t_{i-1})}{\delta t} - \frac{du}{dt}(t_{i-1/2})\Big).\end{aligned}$$

By the above formula we obtain the estimate

$$\delta t \left\| \frac{du}{dt}(t_{i-1/2}) - \frac{u(t_i) - u(t_{i-1})}{\delta t} \right\|_{L^2(U)} \le C(\delta t)^2 \int_{t_{i-1}}^{t_i} \|\tfrac{d^3u}{ds^3}\|_{L^2(U)} ds.$$

By the Taylor formula, we get for the last term

$$\begin{aligned}\frac{1}{2}\big(\tfrac{du}{dt}(t_i) + \tfrac{du}{dt}(t_{i-1})\big) - \tfrac{du}{dt}(t_{i-1/2}) \\ = \frac{1}{2}\int_{t_{i-1}}^{t_{i-1/2}} (s - t_{i-1}) \tfrac{d^3u(s)}{ds^3} ds + \frac{1}{2}\int_{t_{i-1/2}}^{t_i} (t_i - s) \tfrac{d^3u(s)}{ds^3} ds,\end{aligned}$$

which gives the estimate

$$\Big\| \frac{1}{2}\big(\tfrac{du}{dt}(t_i) + \tfrac{du}{dt}(t_{i-1})\big) - \tfrac{du}{dt}(t_{i-1/2}) \Big\|_{L^2(U)} \leq C\delta t \int_{t_{i-1}}^{t_i} \|\tfrac{d^3u}{ds^3}\|_{L^2(U)} ds.$$

Together these estimates give

$$\begin{aligned} \|D_h^n\|_{L^2(U)} &\leq C\, h^2 \|g\|_{H^2(U)} + C\, h^2 \int_0^{t_n} \|\tfrac{du}{ds}\|_{H^2(U)} ds \\ &\qquad + C\,(\delta t)^2 \sum_{i=1}^{n} \int_{t_{i-1}}^{t_i} \|\tfrac{d^3u}{ds^3}\|_{L^2(U)} ds \\ &\leq C\bigg(h^2 \|g\|_{H^2(U)} + h^2 \int_0^{t_n} \|\tfrac{du}{ds}\|_{H^2(U)} ds + (\delta t)^2 \int_0^{t_n} \|\tfrac{d^3u}{ds^3}\|_{L^2(U)} ds \bigg). \end{aligned}$$

■

The error estimates of Theorems 7.22 and 7.23 are given in terms of the exact solution $u(t)$. Using the regularity estimates like in Theorem 5.27 we can estimate the error in terms of the initial data g and the nonhomogeneous term f.

Chapter 8

American options

An American option is a contract that grants the holder the right to buy or sell a security (called the underlying) at an agreed-upon price during some period of time up to and including its maturity date. The option is Bermudan if it can only be exercised at some discrete, finite set of points in time prior to and including the maturity date. Such contracts are traded in all major financial markets, so identifying efficient techniques for pricing them is a very important problem.

The computation of American option prices is challenging problem, especially when several underlying assets are involved. The mathematical problem to solve is an optimal stopping problem. In diffusion models, this problem is reduced to a variational inequality, which is solved by PDE methods. But in the late 1990s, numerical methods based on Monte-Carlo techniques were introduced. The starting point of these methods is the replacement of the interval of exercise dates with a finite set of dates. This amounts to approximating the American option by the Bermudan option.

In this chapter, we will analyze both mentioned above numerical methods. We will present the Monte Carlo procedure due to Longstaff and Schwartz implementing effectively the dynamic programming principle supplemented with least squares regression on a finite set of functions that approximate conditional expectations. We will also analyze algorithms that approximate variational inequalities. We will limit our presentation to two algorithms: projected successive over-relaxation (PSOR) and penalization.

8.1 Pricing American options

We begin with a collection of results from the theory of optimal stopping in continuous time. In this presentation, we follow the paper by El Karoui [18] where

the reader can find more complete results and relevant proofs. We consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with filtration $\mathbb{F} = (\mathcal{F}_t)_{0 \le t \le T}$ satisfying usual conditions of completeness and right-continuity. In addition, we assume that the σ-field $\mathcal{F}_0$ is trivial.

We denote by $\mathcal{T}_{t,T}$ the set of all stopping times $\tau \in [t, T]$ with respect to filtration $\mathbb{F}$

$$\mathcal{T}_{t,T} = \{\tau\colon\ \mathbb{P}(\tau \in [t, T]) = 1\}, \quad 0 \le t < T < +\infty.$$

DEFINITION. 8.1 *An adapted, right-continuous process $(X_t)_{0 \le t \le T}$ is called* regular *if X_τ is integrable for every $\tau \in \mathcal{T}_{0,T}$, and for every nondecreasing sequence τ_n of stopping times with $\lim_{n\to\infty} \tau_n = \tau$, we have $\lim_{n\to\infty} \mathbb{E}(X_{\tau_n}) = \mathbb{E}(X_\tau)$.*

THEOREM. 8.2 *Let $Z = (Z_t)_{0 \le t \le T}$ be an adapted, right-continuous process satisfying $Z_t \ge 0$ for all $t \in [0, T]$, and $\mathbb{E}(\sup_{0 \le t \le T} Z_t) < +\infty$.*
For $t \in [0, T]$, we define

$$U_t = \operatorname*{ess\,sup}_{\tau \in \mathcal{T}_{t,T}} \mathbb{E}(Z_\tau | \mathcal{F}_t).$$

Then

1. *$(U_t)_{0 \le t \le T}$ is a supermartingale.*
2. *$\mathbb{E}(U_t) = \sup_{\tau \in \mathcal{T}_{t,T}} \mathbb{E}(Z_\tau)$.*
3. *U admits a right-continuous modification.*

DEFINITION. 8.3 *A right-continuous modification of the process U from the above theorem is called the* Snell envelope *of Z.*

DEFINITION. 8.4 *For a process Z a stopping time $\hat{\tau} \in \mathcal{T}_{0,T}$ is* optimal *if $\mathbb{E}(Z_{\hat{\tau}}) = \sup_{\tau \in \mathcal{T}_{0,T}} \mathbb{E}(Z_\tau)$.*

THEOREM. 8.5 *A stopping time $\hat{\tau} \in \mathcal{T}_{0,T}$ is optimal if and only if the following conditions hold*

1. *$U_{\hat{\tau}} = Z_{\hat{\tau}}$ a.s.*
2. *$U_t^{\hat{\tau}} = U_{\hat{\tau} \wedge t}$, $0 \le t \le T$, is a martingale.*

THEOREM. 8.6 *If the process Z is regular, then its Snell envelop U is also regular and*

$$\tau_0 = \inf\{s \in [t,T]\colon\ U_s = Z_s\}$$

is the smallest optimal stopping time in $\mathcal{T}_{t,T}$. Define

$$v(t) = \mathbb{E}(U_t)$$

called the value function of the optimal stopping problem *for the process Z in $\mathcal{T}_{t,T}$. Then $v(t)$ is given by the expression*

$$v(t) = \mathbb{E}(U_t) = \sup_{\tau \in \mathcal{T}_{t,T}} \mathbb{E}(Z_\tau) = \mathbb{E}(Z_{\tau_0}).$$

After this brief review of results concerning optimal stopping, we return to American options. The problem of pricing American options will be considered in a financial market of d risky assets whose prices are given by stochastic processes $S_t = (S_t^1, \ldots, S_t^d)$ in $(\Omega, \mathcal{F}, \mathbb{P})$. We assume that these processes are adapted to filtration $\mathbb{F}$, right-continuous with left limits, strictly positive semimartingales. There is a riskless asset in that market that defines a discount factor β_t.

Since the prices S_t^i are nonnegative we introduce, similarly like in Chapter 5, the new variable $X_t = \ln S_t$. The dynamics of X_t is described by the following stochastic differential equation

$$dX_s = b(s, X_s)ds + \sigma(s, X_s)dW_s. \tag{8.1}$$

We will denote by $X_s^{t,x}$ a solution of (8.1) with an initial condition $X_t = x$.

ASSUMPTION. 8.7 *About the coefficients of* (8.1) *we assume:*

(A1) $b(t,x) \in C_b^1([0,T] \times \mathbb{R}^d)$ *is a vector in $\mathbb{R}^d$, where C_b^k denotes k-times differentiable functions which are bounded together with their derivatives to order k.*

(A2) $\sigma(t,x) \in C_b^1([0,T] \times \mathbb{R}^d)$ *is a $d \times d$ matrix. In addition, the "diagonal" entries $\frac{\partial^2 \sigma_i^j}{\partial x_i \partial x_j}$ of the matrix of second derivatives with respect to x are bounded and Hölder continuous uniformly in $(t,x) \in [0,T] \times \mathbb{R}^d$.*

(A3) Matrix $A = \frac{1}{2}\sigma\sigma^\top = (a_{ij})_{i,j=1}^d$ is positive definite

$$\exists \delta > 0\colon\ \sum_{i,j=1}^d a_{ij}(t,x)\xi_i\xi_j \geq \delta\|\xi\|^2, \quad \forall t \in [0,T],\ x \in \mathbb{R}^d,\ \xi \in \mathbb{R}^d \setminus \{0\}.$$

(A4) The discount factor β_t is deterministic $\beta_t = \exp\Big(-\int_0^t r(s)ds\Big)$, where $r(t)$ is a nonnegative function in $C^1([0,T])$.

DEFINITION. 8.8 *An American option for underlying X_t is a nonnegative, adapted stochastic process $(Z_t)_{0\le t\le T}$, $Z_t = g(t, X_t)$, where $g(t,x)$, called the* option reward*, is a continuous function of (t,x) with exponential growth in x*

$$|g(t,x)| \le C\mathrm{e}^{C|x|}, \quad C > 0.$$

Under the conditions of Assumption 8.7, Z_t is a continuous and regular process and satisfies

$$\mathbb{E}\big(\sup_{t\in[0,T]} Z_t\big) < +\infty.$$

THEOREM. 8.9 *Let $(\Omega, \mathcal{F}, \mathbb{P}^*)$ be a probability space with the risk-neutral probability measure $\mathbb{P}^*$, i.e., a measure equivalent to $\mathbb{P}$ such that processes $\beta_t S_t^i$, $i = 1,\dots,d$, are martingales with respect to $\mathbb{P}^*$. Consider an American option in $(\Omega, \mathcal{F}, \mathbb{P}^*)$ written on a underlying asset $X_t = \ln S_t$ which fulfills the conditions of Assumption 8.7. The price at time t of an American option with payoff process $Z_t = g(t, X_t)$, with the reward function $g(t,x)$ fulfilling the conditions of Definition 8.8, is given by $v(t, X_t)$, where*

$$v(t,x) = \frac{1}{\beta_t} \sup_{\tau\in\mathcal{T}_{t,T}} \mathbb{E}^*\big(\beta_\tau g(\tau, X_\tau^{t,x})\big).$$

Furthermore, there is a stopping time $\hat\tau$ attaining this supremum.

The above defined price is called the "fair price" *or* arbitrage-free price *of the American option.*

The following theorem is due to Jaillet, Lamberton, and Lapeyre [26]. It solves the pricing problem for an American option in terms of the value function of an optimal stopping problem.

THEOREM. 8.10 *Let $X_s^{t,x}$ be a solution of* (8.1) *with the coefficients fulfilling Assumption 8.7 and a deterministic initial condition $X_t = x$. Let $g(t,x)$ be the reward function of Definition 8.8. Define the function*

$$V(t,x) = \frac{1}{\beta_t} \sup_{\tau\in\mathcal{T}_{t,T}} \mathbb{E}\Big(\beta_\tau g\big(\tau, X_\tau^{t,x}\big)\Big) \quad (t,x)\in[0,T]\times\mathbb{R}^d. \tag{8.2}$$

This function is continuous, is the value function of the optimal stopping problem for $\tilde Z_t = \beta_t g(t, X_t)$, and the process $\big(\beta_t V(t, X_t)\big)_{0\le t\le T}$ is the Snell envelope of $\tilde Z_t$. Since the initial condition $X_t = x$ is deterministic then

$$V(t, X_t) = \frac{1}{\beta_t} \operatorname*{ess\,sup}_{\tau\in\mathcal{T}_{t,T}} \mathbb{E}\Big(\beta_\tau g\big(\tau, X_\tau\big)\Big|\mathcal{F}_t\Big).$$

For simplicity, we assume in the rest of this chapter that $\mathbb{P}$ is a risk-neutral measure and processes $\beta_t S_t^i$, $i = 1, \ldots, d$, are martingales with respect to this measure.

COROLLARY. 8.11 *Under the above assumption the discounted asset prices are $\mathbb{P}$-martingales and the value function $V(t,x)$ is the arbitrage-free price of the American option with reward $g(t,x)$.*

8.2 Monte Carlo pricing

The most popular Monte Carlo implementation of the valuation formula from Theorem 8.10 is the Longstaff-Schwartz algorithm [37]. For numerical computations, we replace the continuous optimal stopping problem of Definition 8.4 with the following discrete optimal stopping problem.

In the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ we consider a Markov chain $(X_k)_{k=0}^K$ with values in $\mathbb{R}^d$ and a discrete filtration $(\mathcal{F}_k)_{k=0}^K$ generated by this Markov chain and corresponding to a time discretization $0 = t_0 < t_1 < \cdots < t_K = T$.

Given a nonnegative, adapted, discrete, square integrable stochastic process $(Z_k)_{k=0}^K$ with $Z_k = g(t_k, X_k)$ we want to compute $\sup_{\tau \in \mathcal{T}_{0,K}} Z_\tau$, where $\mathcal{T}_{k,K}$ denotes the set of all stopping times with values in $\{t_k, \ldots, t_K\}$. To simplify computations we assume that the discount factor $\beta_t = 1$ for all $t \in [0,T]$.

Let $U = (U_k)_{k=0}^K$ be the Snell envelope of Z

$$U_k = \operatorname*{ess\,sup}_{\tau \in \mathcal{T}_{k,K}} \mathbb{E}\big(Z_\tau \big| \mathcal{F}_k\big), \quad k = 0, \ldots, K.$$

The discrete optimal stopping problem for Z_k and its Snell envelope U_k is solved by the dynamic programming

$$\begin{aligned} &U_K = Z_K, \\ &U_k = \max\big(Z_k, \mathbb{E}(U_{k+1}|\mathcal{F}_k)\big), \quad k = 0, \ldots, K-1. \end{aligned}$$

Let

$$\tau_k = \min\{j \geq k\colon\ U_j = Z_j\}.$$

Then

$$U_k = \mathbb{E}(Z_{\tau_k}\big|\mathcal{F}_k) = \max\big(Z_k, \mathbb{E}(Z_{\tau_{k+1}}|\mathcal{F}_k)\big). \tag{8.3}$$

The above relation enable us to write the dynamic programming principle in terms of the stopping times τ_k:

$$\begin{aligned} &\tau_K = t_K = T, \\ &\tau_k = t_k \mathbb{1}_{\{Z_k \geq \mathbb{E}(Z_{\tau_{k+1}}|\mathcal{F}_k)\}} + \tau_{k+1}\mathbb{1}_{\{Z_k < \mathbb{E}(Z_{\tau_{k+1}}|\mathcal{F}_k)\}}, \quad k = 0, \ldots, K-1, \end{aligned}$$

and obtain the value function for the discrete stopping problem

$$U_0 = \max\big(Z_0, \mathbb{E}(Z_{\tau_1})\big).$$

As $Z_0 = g(0,x)$ is known this gives an approximation of $V(0,x)$.

The algorithm

Step 1.

Since $(X_k)_{k=0}^K$ is a Markov chain then $\mathbb{E}(Z_{\tau_{k+1}}|\mathcal{F}_k) = \mathbb{E}(Z_{\tau_{k+1}}|X_k)$. The state space of $\mathbb{E}(Z_{\tau_{k+1}}|X_k)$ is $L^2(\mathbb{R}^d, \mu_k)$ where μ_k is the measure generated by the random variable X_k. Let $\big(e_m^k(X_k)\big)_{m\geq 1}$ be a basis in $L^2(\mathbb{R}^d, \mu_k)$. Then

$$\mathbb{E}(Z_{\tau_{k+1}}|X_k) = \sum_{m=1}^{\infty} \lambda_m^k e_m^k(X_k). \tag{8.4}$$

We approximate $\mathbb{E}(Z_{\tau_{k+1}}|X_k)$ in an M-dimensional space $\mathcal{H}_k \subset L^2(\mathbb{R}^d, \mu_k)$ defining an orthogonal projection (for simplicity we assume all spaces $\mathcal{H}_k$ of the same dimension M)

$$\pi_k^M \colon L^2(\Omega, \mathbb{P}) \to \mathcal{H}_k.$$

We construct the sequence of approximate stopping times recursively. Starting with $\tau_K^M = t_K = T$ and assuming that τ_{k+1}^M is known we define

$$\pi_k^M\big(Z_{\tau_{k+1}^M}\big) = F_k^M$$

and the stopping time

$$\tau_k^M = t_k \mathbb{1}_{\{Z_k \geq F_k^M\}} + \tau_{k+1}^M \mathbb{1}_{\{Z_k < F_k^M\}}.$$

The stopping times τ_k^M define the processes $Z_{\tau_k^M}$

$$\begin{aligned} Z_{\tau_K^M} &= Z_K, \\ Z_{\tau_k^M} &= Z_k \mathbb{1}_{\{Z_k \geq F_k^M\}} + Z_{\tau_{k+1}^M} \mathbb{1}_{\{Z_k < F_k^M\}}. \end{aligned}$$

Let $e^k(X_k) = \{e_1^k(X_k), \ldots, e_M^k(X_k)\}$ be a basis in $\mathcal{H}_k$. Then

$$F_k^M = \sum_{m=1}^{M} \lambda_m^{M,k} e_m^k(X_k),$$

where the coefficients $\lambda_m^{M,k}$ are determined as a solution to the minimization problem

$$\lambda^{M,k} = \arg\min_{a\in\mathbb{R}^M} \mathbb{E}\Big(Z_{\tau_{k+1}^M} - \sum_{m=1}^{M} a_m e_m^k(X_k)\Big)^2.$$

This approximation gives the approximate value function

$$U_0^M = \max\big(Z_0, \mathbb{E}(Z_{\tau_1^M})\big).$$

Step 2.
To obtain a working algorithm we have to know the projections π_k^M. This is achieved by a Monte Carlo simulation. To this end, we simulate N independent paths $(X_k^1, \ldots, X_k^N)$ and compute $Z_k^n = g(t_k, X_k^n)$. The values of Z_k^n are used to compute recursively coefficients $\hat{\lambda}_m^{M,k}$ approximating $\lambda_m^{M,k}$, and stopping times $\hat{\tau}_k^M$ approximating τ_k^M.

We define the stopping times $\hat{\tau}_k^M$ on each path separately starting the recursion with $\hat{\tau}_{K,n}^M = t_K = T$, $n = 1, \ldots, N$. Knowing $\hat{\tau}_{k+1,n}^M$ we compute $\hat{\lambda}^{M,k}$

$$\hat{\lambda}^{M,k} = \arg\min_{a\in\mathbb{R}^M} \sum_{n=1}^{N}\Big(Z^n_{\hat{\tau}^M_{k+1,n}} - \sum_{m=1}^{M} a_m e_m^k(X_k^n)\Big)^2.$$

This procedure gives values $\hat{\lambda}_m^{M,k}$. Then for $n = 1, \ldots, N$ we define

$$\hat{F}_{k,n}^M = \sum_{m=1}^{M} \hat{\lambda}_m^{M,k} e_m^k(X_k^n)$$

and the stopping times

$$\hat{\tau}_{k,n}^M = t_k \mathbb{1}_{\{Z_k^n \geq \hat{F}_{k,n}^M\}} + \hat{\tau}_{k+1,n}^M \mathbb{1}_{\{Z_k^n < \hat{F}_{k,n}^M)\}}.$$

Recursively, starting from $\hat{\tau}_{K,n}^M$ we compute all $\hat{\tau}_{k,n}^M$ for $k = K-1, \ldots, 0$ and $n = 1, \ldots, N$. Finally, we obtain the approximation of the Snell envelope at $t_0 = 0$

$$U_0^{M,N} = \max\Big(Z_0, \frac{1}{N}\sum_{n=1}^{N} Z^n_{\hat{\tau}^M_{1,n}}\Big),$$

which gives a numerical approximation of the value function $V(0,x)$.

Remark. 8.1 *The above presentation of the algorithm is useful for convergence analysis. In practical computations, Step 1 is limited to fixing the dimensions of*

$\mathcal{H}_k$ and the choice of basis functions in these spaces. Only Step 2 is fully implemented in computations. The choice of basis functions means the choice of the class of functions universal for all spaces $\mathcal{H}_k$ (Longstaff-Schwartz used the Laguerre polynomials). In Step 2 of the algorithm, we evaluate these basis functions at points X_k^n.

Let us observe that since $e^k(X_k) = (e_1^k(X_k), \dots, e_M^k(X_k))$ is a basis in $\mathcal{H}_k$ the coefficients $\lambda_m^{M,k}$ are uniquely defined by

$$\big(\lambda_m^{M,k}\big)_{m=1}^M = (A^k)^{-1}\mathbb{E}\Big(Z_{\tau_{k+1}^M} e^k(X_k)\Big),$$

where

$$(A^k)_{ij} = \mathbb{E}\big(e_i^k(X_k), e_j^k(X_k)\big)$$

and A^k is invertible by the linear independence of e_m^k, $m = 1, \dots, M$.

Analogously

$$\big(\hat{\lambda}_m^{M,k}\big)_{m=1}^M = (\hat{A}^k)^{-1}\frac{1}{N}\sum_{n=1}^N Z^n_{\hat{\tau}^M_{k+1,n}} e^k(X_k^n),$$

where

$$(\hat{A}^k)_{ij} = \frac{1}{N}\sum_{n=1}^N e_i^k(X_k^n), e_j^k(X_k^n).$$

Convergence

Our goal is now to prove that for M, N going to infinity $U_0^{M,N}$ converge to U_0, the value function of the discrete optimal stopping problem. Our proof will follow the approach of Clément, Lamberton, and Protter [11].

First, for vectors $a^l \in \mathbb{R}^M$ and $x_l \in \mathbb{R}^d$ we define the functions $F_k^M(a^l, x_l) = \sum_{m=1}^M a_m^l e_m^k(x_l)$. Then, for parameters $a = (a^1, \dots, a^{K-1})$, $x = (x_1, \dots, x_K)$ and $z = (z_1, \dots, z_K)$, with $a^k \in \mathbb{R}^M$, $x_k \in \mathbb{R}^d$, and $z_k \in \mathbb{R}$, we define the scalar functions

$$W_K(a, x, z) = z_K,$$
$$W_k(a, x, z) = z_k \mathbb{1}_{\{z_k \geq F_k^M(a^k, x_k)\}} + W_{k+1}(a, x, z)\mathbb{1}_{\{z_k < F_k^M(a^k, x_k)\}}.$$

Clearly

$$W_k(\lambda^M, X, Z) = Z_{\tau_k^M}, \quad \text{where } \lambda^M = (\lambda^{M,1}, \dots, \lambda^{M,K-1}),$$

and

$$W_k(\hat{\lambda}^M, X^n, Z^n) = Z^n_{\hat{\tau}^M_{k,n}}, \quad \text{where } \hat{\lambda}^M = (\hat{\lambda}^{M,1}, \dots, \hat{\lambda}^{M,K-1}).$$

LEMMA. 8.12 *For* $k \in \{1, \ldots, K-1\}$

$$|W_k(a,x,z) - W_k(b,x,z)| \leq \Big(\sum_{i=k}^{K} |z_i|\Big) \sum_{i=k}^{K-1} \mathbb{1}_{\{|z_i - F_i^M(b^i,x_i)| \leq |F_i^M(b^i,x_i) - F_i^M(a^i,x_i)|\}}.$$

Proof. Let $B_k(a) = \{z_k \geq F_k^M(a^k, x_k)\}$, $B_k(b) = \{z_k \geq F_k^M(b^k, x_k)\}$. Since

$$W_k(a,x,z) = z_k \mathbb{1}_{B_k(a)} + \sum_{i=k+1}^{K-1} z_i \mathbb{1}_{B_k^c(a)\ldots B_{i-1}^c(a) B_i(a)} + z_K \mathbb{1}_{B_k^c(a)\ldots B_{K-1}^c(a)}$$

then

$$\begin{aligned} W_k(a,x,z) - W_k(b,x,z) &= z_k(\mathbb{1}_{B_k(a)} - \mathbb{1}_{B_k(b)}) \\ &+ \sum_{i=k+1}^{K-1} z_i(\mathbb{1}_{B_k^c(a)\ldots B_{i-1}^c(a) B_i(a)} - \mathbb{1}_{B_k^c(b)\ldots B_{i-1}^c(b) B_i(b)}) \\ &+ z_K(\mathbb{1}_{B_k^c(a)\ldots B_{K-1}^c(a)} - \mathbb{1}_{B_k^c(b)\ldots B_{K-1}^c(b)}), \end{aligned}$$

where $\mathbb{1}_{A^c} = \mathbb{1} - \mathbb{1}_A$ and $\mathbb{1}_{AB} = \mathbb{1}_A + \mathbb{1}_B - \mathbb{1}_A \mathbb{1}_B$.

Since

$$\begin{aligned} &\big|\mathbb{1}_{B_k^c(a)\ldots B_{i-1}^c(a) B_i(a)} - \mathbb{1}_{B_k^c(b)\ldots B_{i-1}^c(b) B_i(b)}\big| \\ &\leq \sum_{j=k}^{i-1} \big|\mathbb{1}_{B_j^c(a)} - \mathbb{1}_{B_j^c(b)}\big| + \big|\mathbb{1}_{B_i(a)} - \mathbb{1}_{B_i(b)}\big| = \sum_{j=k}^{i} \big|\mathbb{1}_{B_j(a)} - \mathbb{1}_{B_j(b)}\big|, \end{aligned}$$

and

$$\big|\mathbb{1}_{B_k^c(a)\ldots B_{K-1}^c(a)} - \mathbb{1}_{B_k^c(b)\ldots B_{K-1}^c(b)}\big| \leq \sum_{j=k}^{K-1} \big|\mathbb{1}_{B_j(a)} - \mathbb{1}_{B_j(b)}\big|$$

we obtain

$$|W_k(a,x,z) - W_k(b,x,z)| \leq \Big(\sum_{i=k}^{K} |z_i|\Big) \sum_{i=k}^{K-1} \big|\mathbb{1}_{B_i(a)} - \mathbb{1}_{B_i(b)}\big|.$$

As $|\mathbb{1}_A - \mathbb{1}_B| = \mathbb{1}_{(A\setminus B)\cup(B\setminus A)}$ we get

$$\begin{aligned}
&\Big|\mathbb{1}_{B_k(a)} - \mathbb{1}_{B_k(b)}\Big| \\
&\qquad = \mathbb{1}_{\{F_k^M(a^k,x_k)\le z_k<F_k^M(b^k,x_k)\}} + \mathbb{1}_{\{F_k^M(b^k,x_k)\le z_k<F_k^M(a^k,x_k)\}} \\
&\qquad = \mathbb{1}_{\{F_k^M(a^k,x_k)-F_k^M(b^k,x_k)\le z_k-F_k^M(b^k,x_k)<0\}} \\
&\qquad\qquad + \mathbb{1}_{\{0\le z_k-F_k^M(b^k,x_k)<F_k^M(a^k,x_k)-F_k^M(b^k,x_k)\}} \\
&\qquad = \mathbb{1}_{\{0<F_k^M(b^k,x_k)-z_k\le F_k^M(b^k,x_k)-F_k^M(a^k,x_k)\}} \\
&\qquad\qquad + \mathbb{1}_{\{0\le z_k-F_k^M(b^k,x_k)<F_k^M(a^k,x_k)-F_k^M(b^k,x_k)\}} \\
&\qquad \le \mathbb{1}_{\{|z_k-F_k^M(b^k,x_k)|\le|F_k^M(b^k,x_k)-F_k^M(a^k,x_k)|\}}.
\end{aligned}$$

Application of the above estimate concludes the proof. ■

To prove the convergence of $U_0^{M,N}$ to U_0 is not an easy task. First, we keep M constant and consider the limit $N\to\infty$ proving that $U_0^{M,N}$ converge to U_0^M.

LEMMA. 8.13 *Let the simulated paths $X^n = (X_1^n,\dots,X_K^n)$, $n=1,\dots,N$, be independent. We assume that in the finite dimensional space $\mathcal{H}_k$*

(LS) $\mathbb{P}(Z_k = F_k^M) = 0$, $k=1,\dots,K-1$.

Under the above assumptions $(\hat\lambda_m^{M,k})_{m=1}^M$ converge almost surely to $(\lambda_m^{M,k})_{m=1}^M$, $k=1,\dots,K-1$.

Proof. The proof goes by induction in k. For $k=K-1$ the result follows from the strong law of large numbers. Assume that the result is true for $i=k,\dots,K-1$. Then

$$\big(\hat\lambda_m^{M,k-1}\big)_{m=1}^M = (\hat A^{k-1})^{-1}\frac{1}{N}\sum_{n=1}^N Z^n_{\hat\tau^M_{k,n}}\, e^{k-1}(X^n_{k-1}).$$

Since $\frac{1}{N}\sum_{n=1}^N W_k(\hat\lambda^M, X^n, Z^n)e^{k-1}(X^n_{k-1})$ converge by the law of large numbers to $\mathbb{E}\Big(W_k(\hat\lambda^M, X, Z)e^{k-1}(X_{k-1})\Big)$ and by the same reason $\hat A^{k-1}$ converge to A^{k-1} then by the equality

$$Z^n_{\hat\tau^M_{k,n}} = W_k(\hat\lambda^M, X^n, Z^n)$$

it suffices to prove

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^N\Big(W_k(\hat\lambda^M, X^n, Z^n) - W_k(\lambda^M, X^n, Z^n)\Big)e^{k-1}(X^n_{k-1}) = 0.$$

By Lemma 8.12 we get

$$\frac{1}{N}\Big|\sum_{n=1}^{N}\Big(W_k(\hat{\lambda}^M,X^n,Z^n)-(W_k(\lambda^M,X^n,Z^n)\Big)e^{k-1}(X_{k-1}^n)\Big|$$
$$\leq \sum_{n=1}^{N}\|e^{k-1}(X_{k-1}^n)\|\sum_{i=k}^{K}|Z_i^n|$$
$$\times\sum_{i=k}^{K-1}\mathbb{1}_{\{|Z_i^n-F_i^M(\lambda^{M,i},X_i^n)|\leq|F_i^M(\hat{\lambda}^{M,i},X_i^n)-F_i^M(\lambda^{M,i},X_i^n)|\}}.$$

Since $\hat{\lambda}^{M,i}$ converge to $\lambda^{M,i}$ for $i\geq k$ then

$$\limsup_{N\to\infty}\frac{1}{N}\Big|\sum_{n=1}^{N}\Big(W_k(\hat{\lambda}^M,X^n,Z^n)-(W_k(\lambda^M,X^n,Z^n)\Big)e^{k-1}(X_{k-1}^n)\Big|$$
$$\leq\limsup_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\|e^{k-1}(X_{k-1}^n)\|\sum_{i=k}^{K}|Z_i^n|\sum_{i=k}^{K-1}\mathbb{1}_{\{|Z_i^n-F_i^M(\lambda^{M,i},X_i^n)|\leq\epsilon\|e^i(X_i^n)\|\}}$$
$$=\mathbb{E}\Big(\|e^{k-1}(X_{k-1})\|\sum_{i=k}^{K}|Z_i|\sum_{i=k}^{K-1}\mathbb{1}_{\{|Z_i-F_i^M(\lambda^{M,i},X_i)|\leq\epsilon\|e^i(X_i)\|\}}\Big),$$

where in the last equality we have applied the strong law of large numbers. Since $\mathbb{P}(Z_i=F_i^M)=0$ for each i by assumption (LS) then letting ϵ to zero we obtain the desired convergence. ■

Let us remark that assumption (LS) is essential for a correct convergence of the optimal stopping times $\hat{\tau}_n^{M*}$ of Z^n on simulated paths X^n to the optimal stopping time τ^{M*} of Z. It can happen that the values of $\hat{F}_{k,n}^M$, which converge to F_k^M, are always on one side of F_k^M and Z^n flips around Z. Then the optimal stopping times $\hat{\tau}_n^{M*}$ of Z^n will not converge to τ^{M*}. Condition (LS) ensures that such an event has a probability zero.

THEOREM. 8.14 *If under the assumptions of Lemma 8.13 $(\hat{\lambda}_m^{M,k})_{m=1}^M$ converge almost surely to $(\lambda_m^{M,k})_{m=1}^M$ as N tends to infinity then $U_0^{M,N}$ converge to U_0^M in probability.*

Proof. Since Z_0 is deterministic it is enough to prove that $\frac{1}{N}\sum_{n=1}^{N}Z^n_{\hat{\tau}_{1,n}^M}$ converge to $\mathbb{E}(Z_{\tau_1^M})$. As paths X^n are independent then $\frac{1}{N}\sum_{n=1}^{N}\sum_{m=1}^{M}\hat{\lambda}_m^{M,k}e_m^k(X_k^n)$ converge to $\mathbb{E}(Z_{\tau_{k+1}^M}|X_k)$. Thus we have to prove

$$\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\Big(Z^n_{\hat{\tau}_{k,n}^M}-W_k(\lambda^M,X,Z)\Big)=0,$$

where $\lambda^M = (\lambda^{M,1}, \dots, \lambda^{M,K-1})$.

Since $Z^n_{\hat{\tau}^M_{k,n}} = W_k(\hat{\lambda}^M, X^n, Z^n)$ we have

$$\begin{aligned}
&\frac{1}{N}\Big|\sum_{n=1}^{N} W_k(\hat{\lambda}^M, X^n, Z^n) - W_k(\lambda^M, X, Z)\Big| \\
&\quad\le \frac{1}{N}\sum_{n=1}^{N}\Big(|W_k(\hat{\lambda}^M, X^n, Z^n) - W_k(\lambda^M, X^n, Z^n)| \\
&\quad\quad + |W_k(\lambda^M, X^n, Z^n) - W_k(\lambda^M, X, Z)|\Big).
\end{aligned}$$

By Lemma 8.12 we have

$$\begin{aligned}
&\frac{1}{N}\sum_{n=1}^{N}|W_k(\hat{\lambda}^M, X^n, Z^n) - W_k(\lambda^M, X^n, Z^n)| \\
&\quad\le \frac{1}{N}\sum_{n=1}^{N}\sum_{i=k}^{K}|Z^n_i|\sum_{i=k}^{K-1}\mathbb{1}_{\{|Z^n_i - F^M_i(\lambda^{M,i}, X^n_i)| \le |F^M_i(\hat{\lambda}^{M,i}, X^n_i) - F^M_i(\lambda^{M,i}, X^n_i)|\}}.
\end{aligned}$$

As $\hat{\lambda}^M$ converge to λ^M almost surely then for some $\epsilon > 0$ we have

$$\begin{aligned}
&\limsup_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}|W_k(\hat{\lambda}^M, X^n, Z^n) - W_k(\lambda^M, X^n, Z^n)| \\
&\quad\le \limsup_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\sum_{i=k}^{K}|Z^n_i|\sum_{i=k}^{K-1}\mathbb{1}_{\{|Z^n_i - F^M_i(\lambda^{M,i}, X^n_i)| \le \epsilon\}} \\
&\quad= \mathbb{E}\Big(\sum_{i=k}^{K}|Z_i|\sum_{i=k}^{K-1}\mathbb{1}_{\{|Z_i - F^M_i(\lambda^{M,i}, X_i)| \le \epsilon\}},
\end{aligned}$$

where the last equality follows from the strong law of large numbers. Letting ϵ to zero and using assumption (LS) we get for $N \to \infty$

$$\frac{1}{N}\sum_{n=1}^{N}\big(W_k(\hat{\lambda}^M, X^n, Z^n) - W_k(\lambda^M, X^n, Z^n)\big) \to 0.$$

The convergence

$$\frac{1}{N}\sum_{n=1}^{N}\big(W_k(\lambda^M, X^n, Z^n) - W_k(\lambda^M, X, Z)\big) \to 0$$

follows straightforwardly from the law of large numbers. ■

The convergence of U_0^M to U_0 is a consequence of the theorem below and the properties of orthogonal projections in a Hilbert space.

THEOREM. 8.15 *The following convergence holds in* $L^2(\mathbb{R}^d, \mu_k)$

$$\lim_{M\to\infty} \mathbb{E}(Z_{\tau_k^M}|X_k) = \mathbb{E}(Z_{\tau_k}|X_k),$$

for $k = 1, \ldots, K$.

Proof. The proof goes by induction in k. For $k = K$ the result is true since $Z_{\tau_K^M} = Z_{\tau_K} = Z_T$. Assume now that the result is true for $k+1$. We will show that it is also true for k.

From the definition

$$Z_{\tau_k^M} = Z_k \mathbb{1}_{\{Z_k \geq F_k^M\}} + Z_{\tau_{k+1}^M} \mathbb{1}_{\{Z_k < F_k^M\}}$$

we obtain using (8.3)

$$\begin{aligned}\mathbb{E}(Z_{\tau_k^M} - Z_{\tau_k}|X_k) =& \big(Z_k - E(Z_{\tau_{k+1}}|X_k)\big)\big(\mathbb{1}_{\{Z_k \geq F_k^M\}} - \mathbb{1}_{\{Z_k \geq \mathbb{E}(Z_{\tau_{k+1}}|X_k)\}}\big) \\ &+ \mathbb{E}(Z_{\tau_{k+1}^M} - Z_{\tau_{k+1}}|X_k)\mathbb{1}_{\{Z_k < F_k^M\}}.\end{aligned}$$

The second term on the right hand side converges to zero by the inductive assumption. For the first term, using the identity $|\mathbb{1}_A - \mathbb{1}_B| = \mathbb{1}_{(A\setminus B)\cup(B\setminus A)}$, we get similarly like in the proof of Lemma 8.12

$$\begin{aligned}&\big|\big(Z_k - E(Z_{\tau_{k+1}}|X_k)\big)\big(\mathbb{1}_{\{Z_k \geq F_k^M\}} - \mathbb{1}_{\{Z_k \geq \mathbb{E}(Z_{\tau_{k+1}}|X_k)\}}\big)\big| \\ &= \big|Z_k - E(Z_{\tau_{k+1}}|X_k)\big|\,\big|\mathbb{1}_{\{\mathbb{E}(Z_{\tau_{k+1}}|X_k) > Z_k \geq F_k^M\}} - \mathbb{1}_{\{F_k^M > Z_k \geq \mathbb{E}(Z_{\tau_{k+1}}|X_k)\}}\big| \\ &\leq \big|Z_k - E(Z_{\tau_{k+1}}|X_k)\big|\mathbb{1}_{\{|Z_k - \mathbb{E}(Z_{\tau_{k+1}}|X_k) \leq |F_k^M - \mathbb{E}(Z_{\tau_{k+1}}|X_k)|\}} \\ &\leq |F_k^M - \mathbb{E}(Z_{\tau_{k+1}}|X_k)| \\ &\leq \big|F_k^M - \pi_k^M\big(\mathbb{E}(Z_{\tau_{k+1}}|X_k)\big)\big| + \big|\pi_k^M\big(\mathbb{E}(Z_{\tau_{k+1}}|X_k)\big) - \mathbb{E}(Z_{\tau_{k+1}}|X_k)\big|.\end{aligned}$$

Since by definition

$$F_k^M = \pi_k^M\big(Z_{\tau_{k+1}^M}\big) = \pi_k^M\big(\mathbb{E}(Z_{\tau_{k+1}^M}|X_k)\big)$$

then

$$\begin{aligned}&\big|\big(Z_k - E(Z_{\tau_{k+1}}|X_k)\big)\big(\mathbb{1}_{\{Z_k \geq F_k^M\}} - \mathbb{1}_{\{Z_k \geq \mathbb{E}(Z_{\tau_{k+1}}|X_k)\}}\big)\big| \\ &\leq \big|\mathbb{E}(Z_{\tau_{k+1}^M}|X_k) - \mathbb{E}(Z_{\tau_{k+1}}|X_k)\big| + \big|\pi_k^M\big(\mathbb{E}(Z_{\tau_{k+1}}|X_k)\big) - \mathbb{E}(Z_{\tau_{k+1}}|X_k)\big|.\end{aligned}$$

The first term on the right hand side converges to zero by the inductive assumption and the second, by properties of orthogonal projections. ■

Hence we have two independent convergences: $U_0^{M,N} \to U_0^M$ as N goes to infinity, and $U_0^M \to U_0$ for $M \to \infty$. This is not sufficient to get the convergence $U_0^{M,N} \to U_0$. The problem of this last convergence has been discussed in many papers in recent years. There are several results obtained with the help of mathematical techniques that are too advanced for the presentation in these lecture notes (cf.[17], [49], [53], [54]). Some of these papers, in addition to proofs of convergence, give also (complicated) formulas of error estimates.

8.3 Variational inequalities

Optimal stopping problems can be reduced to variational inequalities. Since variational inequalities are a classical mathematical technique used in many problems of mathematical physics, numerical algorithms for solving them are known for a long time. Below, we present briefly that approach to the optimal stopping problem for American options and find the value function of that problem. We will give a theorem that guarantees the existence and uniqueness of solutions and describe the θ-scheme, which gives its finite difference approximation. This approximation is formulated as a *linear complementarity problem* (LCP). The solution of LCP creates a numerical challenge of its own. There is a large number of numerical algorithms treating LCP. We will present a rather old and not very fast approach of *projected successive over-relaxation* (PSOR), which is quite popular due to the simplicity of its computer implementation. We will prove that the approximation generated by PSOR converges to the exact solution of LCP. We will discuss as well another popular and more efficient algorithm of penalization.

We start with a brief presentation of more significant theoretical results on variational inequalities applied to optimal stopping problems. In this presentation, we follow the paper by Jaillet, Lamberton, and Lapeyre [26]. The reader is also advised to consult the book by Bensoussan and Lions [5] where more complete proofs can be found.

Since the stochastic process X_t for which we have constructed the value function in Section 8.1 is the logarithmic price, it takes values in $\mathbb{R}^d$. Then the partial differential equation corresponding to the stochastic equation for X_t will also be defined in $\mathbb{R}^d$. Hence, similarly like in Section 5.4, we will investigate that equation in weighted Sobolev's spaces. We consider Sobolev's spaces with weight functions fulfilling the conditions of Definition 5.33. In addition to spaces $L^2_\rho(\mathbb{R}^d)$ and $H^1_\rho(\mathbb{R}^d)$ defined in Section 5.4, we introduce general weighted Sobolev's spaces

$W^{k,p}_\rho(\mathbb{R}^d)$ of functions u: $\mathbb{R}^d \to \mathbb{R}$ such that u and all weak derivatives $D^\alpha u$, for $|\alpha| \le k$, belong to $L^p_\rho(\mathbb{R}^d)$ with the norm

$$\|u\|_{W^{k,p}_\rho} = \left(\sum_{\alpha\le k}\int_{\mathbb{R}^d} |D^\alpha u(x)|^p \rho(x)dx\right)^{1/p}, \quad 1 \le p < +\infty.$$

In spaces $W^{k,p}_\rho(\mathbb{R}^d)$ we consider the differential operator $\mathcal{A}^t$

$$\mathcal{A}^t u = \sum_{i,j=1}^d -\frac{\partial}{\partial x_i}\left(a_{ij}(t,x)\frac{\partial u}{\partial x_j}\right) + \sum_{i=1}^d \bar{b}_i(t,x)\frac{\partial u}{\partial x_i} + c(t,x)u, \tag{8.5}$$

and the associated bilinear form $B^t_\rho[u,v]$

$$B^t_\rho[u,v] = \int_{\mathbb{R}^d}\left(\sum_{i,j=1}^d a_{ij}(t,x)\frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_j} + \sum_{i=1}^d \hat{b}_i(t,x)\frac{\partial u}{\partial x_i}v + c(t,x)uv\right)\rho(x)dx,$$

where

$$\hat{b}_i(t,x) = \bar{b}_i(t,x) + \sum_{j=1}^d a_{ij}(t,x)\frac{1}{\rho(x)}\frac{\partial \rho(x)}{\partial x_j}.$$

Assuming that ρ fulfills the conditions of Definition 5.33, a_{ij}, b_i, c $(i,j=1,\dots,d)$ are in $L^\infty([0,T)\times\mathbb{R}^d)$ and a_{ij} is symmetric, we can prove, following the lines of the proof of Theorem 5.20, the energy estimates with $\alpha,\beta,\gamma > 0$

$$\begin{aligned} &B^t_\rho[u(t),v(t)] \le \alpha\|u(t)\|_{H^1_\rho}\|v(t)\|_{H^1_\rho},\ u(t),v(t)\in H^1_\rho(\mathbb{R}^d)\ a.e.\ t\in[0,T],\\ &\beta\|u(t)\|^2_{H^1_\rho} \le B^t_\rho[u(t),u(t)] + \gamma\|u(t)\|^2_{L^2_\rho},\ u(t)\in H^1_\rho(\mathbb{R}^d)\ a.e.\ t\in[0,T]. \end{aligned} \tag{8.6}$$

The market model described by (8.1) corresponds to operator $\mathcal{A}^t$ with coefficients

$$a_{ij} = \frac{1}{2}\sum_{k=1}^d \sigma^k_i\sigma^k_j, \quad \bar{b}_i = -b_i + \sum_{j=1}^d \frac{\partial a_{ij}}{\partial x_j}, \quad c = r(t). \tag{8.7}$$

For that specification of $\mathcal{A}^t$ it is easy to prove that under Assumption 8.7 the estimates (8.6) hold.

As we already know from Section 5.4, it is more convenient to operate with a coercive bilinear form. The coerciveness of $B^t_\rho[u,v]$ can be achieved by the change of variables $u_\gamma(t) = \mathrm{e}^{-\gamma t}u(t)$ (see Remark 5.4 in Section 5.4). We assume that this change of variables has been performed and the bilinear form is coercive. To simplify notation, we drop the index γ, which indicates this change. Hence, the differential problem we are dealing with is formulated as follows.

ASSUMPTION. 8.16 *Let $\mathcal{A}^t$ be defined by* (8.5) *and $B^t_\rho[u,v]$ be the corresponding bilinear form. We assume that operator $\frac{\partial}{\partial t}+\mathcal{A}^t$ is uniformly parabolic in the sense of Definition 5.24 and $B^t_\rho[u,v]$ is t-uniformly continuous and coercive in $H^1_\rho(\mathbb{R}^d)$*

$$B^t_\rho[u(t),v(t)] \le \alpha\|u(t)\|_{H^1_\rho}\|v(t)\|_{H^1_\rho},\ u(t),v(t)\in H^1_\rho(\mathbb{R}^d)\ a.e.\ t\in[0,T],$$
$$\beta\|u(t)\|^2_{H^1_\rho} \le B^t_\rho[u(t),u(t)],\ u(t)\in H^1_\rho(\mathbb{R}^d)\ a.e.\ t\in[0,T].$$

THEOREM. 8.17 *Assume that B^t_ρ fulfills the conditions of Assumption 8.16 and $g\in H^1_\rho(\mathbb{R}^d)$. Then for each $v\in H^1_\rho(\mathbb{R}^d)$ such that $v\ge g$, there exists a unique solution u defined on $[0,T]\times\mathbb{R}^d$ of the following variational inequality*

$$\begin{aligned} &-\Big(\frac{du}{dt}(t),v-u(t)\Big)_{L^2_\rho}+B^t_\rho[u(t),v-u(t)]\ge 0,\ a.e.\ t\in[0,T),\\ &u(t)\ge g,\quad a.e.\ t\in[0,T],\\ &u(T)=g. \end{aligned} \tag{8.8}$$

For this solution we have $u\in L^2(0,T;H^1_\rho(\mathbb{R}^d))$ and $\frac{du}{dt}\in L^2(0,T;H^{-1}_\rho(\mathbb{R}^d))$.

Remark. 8.2 *From Theorem 5.29 we know that in fact $\frac{du}{dt}\in L^2(0,T;L^2_\rho(\mathbb{R}^d))$ and due to Theorem 5.16 $u\in C(0,T;L^2_\rho(\mathbb{R}^d))$. Hence it is legitimate to use the scalar product $\big(\frac{du}{dt}(t),v-u(t)\big)_{L^2_\rho}$. For the same reason $u(T)=g$ has the standard sense of equality between two L^2_ρ functions.*

THEOREM. 8.18 *Let $g=g(x)\in W^{1,p}_\rho(\mathbb{R}^d)$ with $p>d$. If the hypotheses of Theorem 8.10 are fulfilled and $V(t,x)$ is the value function defined in this theorem then $V(t,x)$ is a solution of the variational inequality* (8.8) *with the coefficients of $B^t_\rho[u,v]$ given by* (8.7)*. (In fact,* (8.8) *is solved by $\mathrm{e}^{-\gamma t}V(t,x)$ due to the mentioned earlier change of variables.)*

The proofs of the above two theorems can be found in Chapter 3 of the book by Bensoussan and Lions [5].

THEOREM. 8.19 *Under the assumptions of Theorem 8.17 variational inequality* (8.8) *is equivalent to the following integral version which holds for each $v\in L^2(0,T;H^1_\rho(\mathbb{R}^d))$ such that $\frac{dv}{dt}\in L^2(0,T;H^{-1}_\rho(\mathbb{R}^d))$ and $v(t)\ge g$, for almost each $t\in[0,T]$,*

$$\begin{aligned} &\int_0^T -\Big(\frac{du}{dt}(t),v(t)-u(t)\Big)_{L^2_\rho}dt+\int_0^T B^t_\rho[u(t),v(t)-u(t)]dt\ge 0,\\ &u(t)\ge g,\quad a.e.\ t\in[0,T],\\ &u(T)=g. \end{aligned} \tag{8.9}$$

In addition, for such v we have

$$\begin{aligned}
&\int_0^T -\Big\langle \frac{dv}{dt}(t), v(t)-u(t)\Big\rangle dt + \int_0^T B_\rho^t[u(t), v(t)-u(t)]dt \\
&\geq \frac{1}{2}\|v(0)-u(0)\|^2_{L^2_\rho} - \frac{1}{2}\|v(T)-u(T)\|^2_{L^2_\rho}, \\
&u(t) \geq g, \quad a.e.\ t \in [0,T], \\
&u(T) = g.
\end{aligned} \tag{8.10}$$

Proof. The fact that (8.8) implies (8.9) follows by integration and properties of functions $u(t)$ and $v(t)$. The opposite implication is due to the fact that $v(t)$ is an arbitrary L^2 function of $t \in [0,T]$.

To obtain (8.10) we have to add to (8.9) the obvious identity

$$\begin{aligned}
&\int_0^T \Big\langle \frac{du}{dt}(t) - \frac{dv}{dt}(t), v(t)-u(t)\Big\rangle dt \\
&= \frac{1}{2}\|v(0)-u(0)\|^2_{L^2_\rho} - \frac{1}{2}\|v(T)-u(T)\|^2_{L^2_\rho}.
\end{aligned}$$

■

DEFINITION. 8.20 *The following problem is called the* weak formulation of variational inequality*:*
find $u \in L^2(0,T;H^1_\rho(\mathbb{R}^d))$ such that $\frac{du}{dt} \in L^2(0,T;H^{-1}_\rho(\mathbb{R}^d))$ and $u(t) \geq g$, for a.e. $t \in [0,T]$, which fulfills the following inequality

$$\begin{aligned}
&\int_0^T -\Big\langle \frac{dv}{dt}(t), v(t)-u(t)\Big\rangle dt + \int_0^T B_\rho^t[u(t), v(t)-u(t)]dt \\
&\geq -\frac{1}{2}\|v(T)-u(T)\|^2_{L^2_\rho},
\end{aligned} \tag{8.11}$$

for each $v \in L^2(0,T;H^1_\rho(\mathbb{R}^d))$ such that $\frac{dv}{dt} \in L^2(0,T;H^{-1}_\rho(\mathbb{R}^d))$ and $v(t) \geq g$, for a.e. $t \in [0,T]$.

Theorem 8.19 shows that a solution of (8.8) is also a solution of the weak formulation (8.11). The following theorem due to Brésis [9] gives the conditions for the opposite implication.

THEOREM. 8.21 *If the bilinear form $B^t_\rho[u,v]$ is t-uniformly continuous and coercive in $H^1_\rho(\mathbb{R}^d)$ the weak formulation* (8.11) *possesses a unique solution u which is also a solution of variational inequality* (8.8)*.*

Remark. 8.3 *In general, the weak formulation of variational inequality possesses many solutions (the advantage of the weak formulation is that a solution always exists). Due to special properties of B^t_ρ, there is only one solution of the weak formulation, which is then also a solution of* (8.8).

For numerical applications, it is convenient to formulate variational inequalities as linear complementarity problems.

LEMMA. 8.22 *Under the assumptions of Theorem 8.17 a solution of* (8.8) *is equivalent to a solution of the following linear complementarity problem*

$$
\begin{aligned}
&i)\ -\Big(\frac{du}{dt}(t), w\Big)_{L^2_\rho} + B^t_\rho[u(t), w] \geq 0, a.e.\ t \in [0,T), w \in H^1_\rho(\mathbb{R}^d), w \geq 0,\\
&ii)\ u(t) \geq g,\ a.e.\ t \in [0,T],\\
&iii)\ -\Big(\frac{du}{dt}(t), g-u(t)\Big)_{L^2_\rho} + B^t_\rho[u(t), g-u(t)] = 0,\ a.e.\ t \in [0,T),\\
&iv)\ u(T) = g.
\end{aligned}
$$

Proof. Let u be a solution of Theorem 8.17. For $w \in H^1_\rho(\mathbb{R}^d)$, $w \geq 0$, we put $v = u + w$. Then $v(t) \geq g$, $v(t) \in H^1_\rho(\mathbb{R}^d)$, and due to (8.8)

$$
\begin{aligned}
&-\Big(\frac{du}{dt}(t), w\Big)_{L^2_\rho} + B^t_\rho[u(t), w]\\
&\quad = -\Big(\frac{du}{dt}(t), v(t)-u(t)\Big)_{L^2_\rho} + B^t_\rho[u(t), v(t)-u(t)] \geq 0,
\end{aligned}
$$

which proves $i)$.

Taking $v(t) = g$ in the above inequality, we obtain

$$
-\Big(\frac{du}{dt}(t), g-u(t)\Big)_{L^2_\rho} + B^t_\rho[u(t), g-u(t)] \geq 0.
$$

On the other hand, taking $w = u(t) - g \geq 0$ in $i)$ we obtain the opposite inequality. That gives

$$
-\Big(\frac{du}{dt}(t), g-u(t)\Big)_{L^2_\rho} + B^t_\rho[u(t), g-u(t)] = 0,
$$

which proves $iii)$.

Conversely, let u solve the LCP of the lemma. Then any $v \in H^1_\rho(\mathbb{R}^d)$, $v \geq g$, can be written as $v = g + w$ with $w \geq 0$ and $w \in H^1_\rho(\mathbb{R}^d)$. By inequality $i)$ we get

$$
-\Big(\frac{du}{dt}(t), v-g\Big)_{L^2_\rho} + B^t_\rho[u(t), v-g] \geq 0.
$$

Adding this inequality to $iii)$ we obtain inequality (8.8).

Since points $ii)$ and $iv)$ are the same as in Theorem 8.17 that completes the proof. ■

To compute a numerical approximation of variational inequality (8.8) we have to restrict our considerations to a compact subset of $\mathbb{R}^d$. We will show how to construct the approximation that converges to a solution to the original problem.

To simplify considerations, we restrict the analysis to the stochastic differential equation with time-independent coefficients

$$dX_s = b(X_s)ds + \sigma(X_s)dW_s$$

and initial condition $X_t = x$.

For numerical computations, we replace the value function of the optimal stopping problem for an American option

$$V(t,x) = \frac{1}{\beta_t} \sup_{\tau \in \mathcal{T}_{t,T}} \mathbb{E}\Big(\beta_\tau g\big(X_\tau^{t,x}\big)\Big).$$

by the approximate value function

$$V_K(t,x) = \frac{1}{\beta_t} \sup_{\tau \in \mathcal{T}_{t,T}} \mathbb{E}\Big(\beta_{\tau \wedge T_K^{t,x}}\, g\Big(X^{t,x}_{\tau \wedge T_K^{t,x}}\Big)\Big),$$

where $T_K^{t,x} = \inf\{s > t\colon\ |X_s^{t,x}| > K\}$.

Then we have the uniform convergence of V_K to V on compact sets (cf. [26] and [55]).

THEOREM. 8.23 *Let functions b, σ and r fulfill Assumption 8.7 and the reward function g fulfill the condition*

(A5) *$\rho(x)g(x)$ is a continuous function on $\mathbb{R}^d$ such that $\|\rho g\|_{L^\infty(\mathbb{R}^d)} \le C$ and $\|D(\rho g)\|_{L^\infty(\mathbb{R}^d)} \le C$ for $C > 0$, where ρ is the weight function of Definition 5.33.*

Then for all $U_R = \{x \in \mathbb{R}^d\colon\ |x| < R\}$, $R > 0$,

$$\lim_{K\to\infty} \max_{t\in[0,T]} \|V(t,\cdot) - V_K(t,\cdot)\|_{L^\infty(\bar{U}_R)} = 0.$$

THEOREM. 8.24 *Let us consider the bilinear form $B_R[u,v]$ defined for $u, v \in H^1(U_R)$ by the formula*

$$B_R[u,v] = \int_{U_R} \Bigg(\sum_{i,j=1}^d a_{ij}(x)\frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_j} + \sum_{i=1}^d \bar{b}_i(x)\frac{\partial u}{\partial x_i}v + c(x)uv\Bigg)dx,$$

where the coefficients of B_R are defined by (8.7)*. If conditions (A1)–(A5) are fulfilled the value function $V_R(t,x)$ is a unique solution of the following variational inequality*

$$\begin{aligned} &-\int_0^T \Big(\frac{du}{dt}(t), v(t)-u(t)\Big)_{L^2(U_R)} dt + \int_0^T B_R[u(t), v(t)-u(t)]dt \geq 0, \\ &u(t) \geq g, \quad a.e.\ t \in [0,T], \\ &u(T) = g, \\ &u(t) = g, \ \ on\ \ \partial U_R, t \in [0,T], \end{aligned} \tag{8.12}$$

which holds for each $v \in L^2(0,T;H^1(U_R))$ such that $\frac{dv}{dt} \in L^2(0,T;H^{-1}(U_R))$ and $v(t) \geq g$ for almost all $t \in [0,T]$, with $u \in L^2(0,T;H^1(U_R))$ and $\frac{du}{dt} \in L^2(0,T;L^2(U_R))$.

Remark. 8.4 *Restricting considerations to a compact set U_R, we have to complement the problem with boundary conditions on ∂U_R. In general, we can impose $u(t) = g_0$ on ∂U_R, where g_0 is an arbitrary function such that $g_0 \geq g$. The result of Theorem 8.23 justifies any boundary conditions. It says that the behavior of the solution near the distant boundary ∂U_R does not affect the solution on any fixed bounded region in the limit $R \to \infty$. Therefore, any well-posed problem on U_R is suitable as an approximation of the original problem on $\mathbb{R}^d$, provided that R is taken sufficiently large. The choice $g_0 \equiv g$ is just the simplest possible choice.*

Discrete variational inequalities

We will now construct a numerical algorithm approximating variational inequality (8.12). A large part of this section follows the ideas of the book by Glowinski, Lions, and Trémollières [22], but the proofs are slightly different as we use finite differences instead of finite elements used in [22]. For simplicity, we will consider only a one-dimensional problem with constant coefficients. Without loss of generality we can assume $U_R = (-1,1)$ with boundary conditions $u(t,-1) = g(-1)$, $u(t,1) = g(1)$. We consider the bilinear form

$$B_R[u,v] = \int_{U_R} \left(a^2 \frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + b\frac{\partial u}{\partial x}v + cuv \right) dx, \tag{8.13}$$

generated by the operator

$$\mathcal{A}u = -a^2 \frac{\partial^2 u}{\partial x^2} + b\frac{\partial u}{\partial x} + cu. \tag{8.14}$$

Under the conditions discussed at the beginning of this Section we have

$$B_R[u,u] \geq \beta \|u\|^2_{H^1(U_R)},$$
$$B_R[u,v] \leq \alpha \|u\|_{H^1(U_R)} \|v\|_{H^1(U_R)}.$$

We approximate operator $-\frac{\partial}{\partial t} + \mathcal{A}$ by finite differences in time and space. We divide $[-1,1]$ into M subintervals with length $\delta x = \frac{2}{M}$ and the time interval $[0,T]$ into N subintervals with length $\delta t = \frac{T}{N}$. With grid points $x_k = -1 + k\delta x$, $k = 0,\ldots,M$ we define a grid function $v_M = (v_{(0)},\ldots,v_{(M)})$ which is an $(M+1)$ dimensional vector. Similarly, taking $t_n = n\delta t$, $n = 0,\ldots,N$, we define a grid function $v_{N,M} = (v_M^{(0)},\ldots,v_M^{(N)})$ in both time and spatial variables. We will write v_M for a vector and $v_{(k)}$ for its components if we consider a grid function in a spatial variable only. Notation $v_{N,M}$ and $v_{(k)}^{(n)}$ will be used for a function and its entries in the case of a grid function in both time and spatial variables.

To define a finite difference approximation of $-\frac{\partial}{\partial t} + \mathcal{A}$ we approximate derivatives by finite differences. The finite difference operators in the x directions are extensions of the operators defined in Section 6.6. The first-order operator δ_x is acting on a grid function v_M by the central differences

$$(\delta_x v_M)_k = \frac{v_{(k+1)} - v_{(k-1)}}{2\delta x}$$

and similarly, the second-order operator

$$(\delta_{xx} v_M)_k = \frac{v_{(k+1)} - 2v_{(k)} + v_{(k-1)}}{(\delta x)^2}.$$

The derivative with respect to time is approximated by the forward difference

$$(\delta_t v_{N,M})^{(n)} = \frac{v_M^{(n+1)} - v_M^{(n)}}{\delta t}.$$

DEFINITION. 8.25 *With the above definitions of the finite difference operators, we define the following spaces for grid functions in spatial variables:*

$$L^2_M(U_R) = \{v_M = (v_{(k)})_{k=0}^M \colon \sum_{k=0}^M |v_{(k)}|^2 < +\infty\}$$

with the norm

$$\|v_M\|^2_{L^2_M} = \sum_{k=0}^M \delta x |v_{(k)}|^2;$$

and

$$H^1_M(U_R) = \{v_M = (v_{(k)})_{k=0}^M \colon \sum_{k=0}^M \Big| \frac{v_{(k+1)} - v_{(k-1)}}{2\delta x} \Big|^2 < +\infty\}$$

with the norm

$$\|v_M\|^2_{H^1_M} = \sum_{k=0}^M \delta x |v_{(k)}|^2 + \sum_{k=0}^M \delta x \Big| \frac{v_{(k+1)} - v_{(k-1)}}{2\delta x} \Big|^2.$$

Similarly, we define the L^2 type spaces for grid functions in time and spatial variables

$$L^2_N(0,T; L^2_M(U_R)) = \{v_{N,M} = (v^{(n)}_M)_{n=0}^N \colon \sum_{n=0}^N \sum_{k=0}^M |v^{(n)}_{(k)}|^2 < +\infty\}$$

with the norm

$$\|v_{N,M}\|^2_{L^2_N(L^2_M)} = \sum_{n=0}^N \sum_{k=0}^M \delta t \delta x |v^{(n)}_{(k)}|^2,$$

and

$$L^2_N(0,T; H^1_M(U_R)) = \{v_{N,M} = (v^{(n)}_M)_{n=0}^N \colon \sum_{n=0}^N \sum_{k=0}^M \Big| \frac{v^{(n)}_{(k+1)} - v^{(n)}_{(k-1)}}{2\delta x} \Big|^2 < +\infty\}$$

with the norm

$$\|v_{N,M}\|^2_{L^2_N(H^1_M)} = \sum_{n=0}^N \sum_{k=0}^M \delta t \delta x |v^{(n)}_{(k)}|^2 + \sum_{n=0}^N \sum_{k=0}^M \delta t \delta x \Big| \frac{v^{(n)}_{(k+1)} - v^{(n)}_{(k-1)}}{2\delta x} \Big|^2,$$

We now define a matrix Λ which is a discrete approximation of operator $\mathcal{A}$.

DEFINITION. 8.26 *Let Λ be an $(M+1) \times (M+1)$ dimensional matrix acting on grid functions v_M*

$$(\Lambda v_M)_{(k)} = -a^2(\delta_{xx} v_M)_{(k)} + b(\delta_x v_M)_{(k)} + c(v_M)_{(k)},\ k = 1, \ldots, M-1,$$
$$(\Lambda v_M)_{(0)} = v_{(0)}, \quad (\Lambda v_M)_{(M)} = v_{(M)}.$$

Then for $v_M, w_M \in H^1_M(U_R)$, we define

$$(\Lambda v_M, w_M)_{L^2_M} = \sum_{k=0}^M \delta x w_{(k)} (\Lambda v_M)_{(k)}$$

as an approximation of the bilinear form $B_R[v,w]$. *By elementary computations we obtain the analogous estimates as for* $B_R[v,w]$

$$|(\Lambda v_M, w_M)_{L^2_M}| \le \alpha \|v_M\|_{H^1_M} \|w_M\|_{H^1_M},$$
$$|(\Lambda v_M, v_M)_{L^2_M}| \ge \beta \|v_M\|^2_{H^1_M}.$$

Our goal is to prove that the grid functions obtained by solving an appropriate discrete variational inequality converge to a solution of the continuous variational inequality (8.12). This requires an extension of grid functions to functions defined on $[0,T] \times U_R$. For this purpose, we use a piecewise constant extension.

Let $\chi(x_k)$ denote the characteristic function of the interval $(x_k - \frac{1}{2}\delta x, x_k + \frac{1}{2}\delta x)$ and $\chi(t_n)$, the characteristic function of the interval $(t_n - \frac{1}{2}\delta t, t_n + \frac{1}{2}\delta t)$, with obvious modifications for $\chi(x_0)$, $\chi(x_M)$, $\chi(t_0)$, $\chi(t_N)$. Then we define the step function

$$\hat{v}_{N,M} = \sum_{n=0}^{N} \sum_{k=0}^{M} v^{(n)}_{(k)} \chi(x_k)\chi(t_n) \tag{8.15}$$

and similarly

$$\hat{v}^{(n)}_M = \sum_{k=0}^{M} v^{(n)}_{(k)} \chi(x_k).$$

The following technical lemma summarizes the essential properties of these step functions.

LEMMA. 8.27 *Let* $v \in L^2(0,T;H^1(U_R))$ *and* $\frac{dv}{dt} \in L^2(0,T;H^{-1}(U_R))$. *For a family of grids defined by uniform partitions of* U_R *with steps* $\delta x \to 0$ *and* $[0,T]$ *with steps* $\delta t \to 0$ *there exists a sequence of grid functions* $v_{N,M}$ *such that the corresponding piecewise constant extensions* $\hat{v}_{N,M}$ *converge for* $\delta x, \delta t \to 0$

$$\hat{v}_{N,M} \to v \text{ in } L^2(0,T;L^2(U_R)),$$
$$(\delta_x v_{N,M})\hat{} \to Dv \text{ in } L^2(0,T;L^2(U_R)),$$
$$(\delta_t v_{N,M})\hat{} \to \tfrac{dv}{dt} \text{ in } L^2(0,T;H^{-1}(U_R)).$$

Here we write Dv *(despite the one-dimensionality of* v*) to stress the fact that the derivative is in the weak sense.*

Proof. Assuming $v \in H^1(U_R)$ we will prove that there is a sequence of grid functions v_M such that $\hat{v}_M \to v$ in $L^2(U_R)$ and $(\delta_x v_M)\hat{} \to Dv$ in $L^2(U_R)$.

Since a function in $H^1(U_R)$ can be approximated by a sequence of C^∞ functions (cf. Theorem 5.6) we can restrict the proof to $v \in C^\infty(\bar{U}_R)$. For a C^∞ function v a weak derivative is a classical derivative which will be denoted by v'.

Let x_k, $k = 0, \ldots, M$, be grid points in U_R. We define the grid function $v_M = (v_{(0)}, \ldots, v_{(M)})$ by the localization of v

$$v_{(k)} = v(x_k).$$

Then there is a sequence of grid functions v_M such that step functions $\hat{v}_M$ converge uniformly to v on U_R. Since v is a continuous function on a compact set, this function is uniformly continuous. Hence, for a given $\epsilon > 0$ there exists a grid step $\delta x = h_1$ such that

$$|v(x) - v(x_k)| < \epsilon \quad \text{for } |x - x_k| < h_1.$$

Then

$$\sup_{U_R} |\hat{v}_M - v| < \epsilon$$

for a grid step h_1. This proves the uniform convergence of $\hat{v}_M$ to v on U_R.

Next, we prove that there exists a grid step h_2 such that $(\delta_x v_M)\hat{}$ converge uniformly to v' on U_R.

With the Taylor expansion, we have

$$\Big|\frac{v(x_k + h_2) - v(x_k - h_2)}{2h_2} - v'(x_k)\Big| \le \frac{1}{2} h_2^2 |v''(x_k + \theta h_2)| \le C h_2^2,$$

where $C = \frac{1}{2} \sup_{U_R} |v''|$.

Let v'_M denote the localization of v'. Then the above Taylor estimate gives

$$\sup_{U_R} \big|(\delta_x v_M)\hat{} - \hat{v}'_M\big| \le C h_2^2 < \epsilon/2$$

for $h_2^2 < \frac{\epsilon}{2C}$.

Since v' is uniformly continuous on U_R then for h_2 sufficiently small we have

$$\sup_{U_R} |\hat{v}'_M - v'| < \epsilon/2.$$

Together these estimates prove the uniform convergence of $(\delta_x v_M)\hat{}$ to v' on U_R.

The proof of convergence $(\delta_t v_{N,M})\hat{} \to \frac{dv}{dt}$ is analogous and will be omitted. ■

Remark. 8.5 *We cannot get $\hat{v}_{N,M} \to v$ in $L^2(0, T; H^1(U_R))$ since $\hat{v}_{N,M}$ is not in $H^1(U_R)$. That is the reason why we consider separately the convergence of $\hat{v}_{N,M}$ to v and $(\delta_x v_{N,M})\hat{}$ to Dv.*

The consistency of finite difference approximations requires a slightly different definition than in Chapter 6.

DEFINITION. 8.28 *We say that a bilinear form (Av_M, w_M) defined by matrix A is consistent with a bilinear form $B[v,w]$ associated with a second-order uniformly elliptic operator $\mathcal{A}$ in divergence form if for each $v_M, w_M \in H^1_M$ such that for $\delta x \to 0$*

$$\hat{v}_M \rightharpoonup v, \quad (\delta_x v_M)\hat{} \rightharpoonup Dv \text{ weakly in } L^2(U_R),$$
$$\hat{w}_M \to w, \quad (\delta_x w_M)\hat{} \to Dw \text{ strongly in } L^2(U_R),$$

we have

$$(A\hat{v}_M, \hat{w}_M) \to B[v,w],$$
$$(A\hat{w}_M, \hat{v}_M) \to B[w,v].$$

With a certain abuse of notation, we write $A\hat{v}_M$ understanding this expression as the abbreviation of $(Av_M)\hat{}$.

LEMMA. 8.29 *The bilinear form $(\Lambda v_M, w_M)$ defined by matrix Λ of Definition 8.26 is consistent in the sense of the above definition with the bilinear form $B_R[v,w]$ given by* (8.13).

In addition, for $v_{N,M} \in L^2_N(0,T; H^1_M(U_R))$ such that for $\delta t, \delta x \to 0$

$$\hat{v}_{N,M} \rightharpoonup v, \quad (\delta_x v_{N,M})\hat{} \rightharpoonup Dv \text{ weakly in } L^2(0,T; L^2(U_R)),$$

we have

$$\liminf_{\delta t, \delta x \to 0} \int_0^T (\Lambda \hat{v}_{N,M}, \hat{v}_{N,M})_{L^2(U_R)} dt \geq \int_0^T B_R[v,v] dt. \tag{8.16}$$

Proof. By elementary computations we get (with natural modifications for $k=0$ and $k=M$)

$$-(\delta_{xx} v_M) w_M = -\sum_{k=0}^{M} \frac{1}{\delta x}\Big(\frac{v_{(k+1)} - v_{(k)}}{\delta x} - \frac{v_{(k)} - v_{(k-1)}}{\delta x}\Big) w_{(k)}$$
$$= \sum_{k=0}^{M} \frac{v_{(k)} - v_{(k-1)}}{\delta x} \frac{w_{(k)} - w_{(k-1)}}{\delta x}.$$

The above identity gives

$$\int_{U_R} -(\delta_{xx}\hat{v}_M)\hat{w}_M dx = \int_{U_R} (\delta_x \hat{v}_M)(\delta_x \hat{w}_M) dx,$$

which enables the rewriting of $(\Lambda \hat{v}_M, \hat{w}_M)$ as the bilinear form

$$(\Lambda \hat{v}_M, \hat{w}_M) = \int_{U_R} \Big(a^2 (\delta_x \hat{v}_M)(\delta_x \hat{w}_M) + b(\delta_x \hat{v}_M)\hat{w}_M + c\hat{v}_M \hat{w}_M \Big) dx.$$

Let now $\hat{v}_M$ and $\hat{w}_M$ be sequences that fulfill the conditions of Definition 8.28. Then $(\Lambda\hat{v}_M, \hat{w}_M)$ and $(\Lambda\hat{w}_M, \hat{v}_M)$ can be expressed in terms of the above bilinear form as linear combinations of scalar products (u_n^1, u_m^2) in $L^2(U_R)$, where u_n^1 is a sequence weakly convergent to u^1 and u_m^2, a sequence strongly convergent to u^2. From the definitions of weak and strong convergence, we have

$$(u_n^1, u_m^2) \to (u^1, u_m^2) \to (u^1, u^2).$$

This proves the consistency of $(\Lambda v_m, w_M)$ with $B_R[v, w]$.

The proof of (8.16) follows from the definition of weak convergence and the Fatou lemma

$$\liminf_{\delta t, \delta x \to 0} \int_0^T \Big((\Lambda\hat{v}_{N,M}, \hat{v}_{N,M}) - (\Lambda\hat{v}_{N,M}, v) \Big) dt$$
$$\geq \liminf_{\delta t, \delta x \to 0} \int_0^T \Big((\Lambda v, \hat{v}_{N,M}) - B_R[v, v] \Big) dt.$$

Then

$$\liminf_{\delta t, \delta x \to 0} \int_0^T (\Lambda\hat{v}_{N,M}, \hat{v}_{N,M}) dt$$
$$\geq \liminf_{\delta t, \delta x \to 0} \int_0^T \Big((\Lambda\hat{v}_{N,M}, v) + (\Lambda v, \hat{v}_{N,M}) - B_R[v, v] \Big) dt \geq \int_0^T B_R[v, v] dt,$$

which ends the proof. ■

We approximate the variational inequality (8.12) by the θ-scheme in time and finite differences in space. Using finite differences in spatial variables creates additional difficulties in the proof of convergence as $\hat{v}_M$ is not in $H^1(U_R)$. In one dimension, one can omit that difficulty by using finite elements from $H^1(U_R)$. We have decided to carry on the proof for finite differences but in a way that can be extended straightforwardly to the finite element approximation in a multidimensional case when $\hat{v}_M$ is not in $H^1(U_R)$. In this way, we obtain simultaneously the proof for finite differences and finite elements.

DEFINITION. 8.30 *Let $g_h = (g_{(0)}, \ldots, g_{(M)})$, $g_{(k)} = g(x_k)$, be an approximation of the reward function g. The* finite difference approximation *$w_{N,M}$ of the variational inequality* (8.12) *is defined recursively by the following θ-scheme:*

i) $w_M^{(N)} = g_h$,

ii) knowing $w_M^{(n+1)} \in H_M^1(U_R)$, $w_M^{(n+1)} \geq g_h$, *find* $w_M^{(n)} \in H_M^1(U_R)$, $w_M^{(n)} \geq g_h$, $n = N-1, \ldots, 0$, *such that for all* $v_M \in H_M^1(U_R)$, $v_M \geq g_h$,

$$\Big(w_M^{(n+1)} - w_M^{(n)} - \delta t \Lambda\big((1-\theta) w_M^{(n+1)} + \theta w_M^{(n)}\big), v_M - w_M^{(n)} \Big)_{L_M^2} \leq 0. \quad (8.17)$$

Since the iterations go downwards then $\theta = 1$ corresponds to the fully implicit scheme.

Let us assume for a moment that the recursion of the above definition possesses a unique solution. The existence of that solution will be proved by the construction of a relevant numerical algorithm (cf. Theorems 8.35 and 8.36).

Our goal now is to prove that $w_{N,M}$ given by (8.17) converges to a solution of (8.12). We begin with the proof of stability for the finite difference approximation.

THEOREM. 8.31 *Let operator Λ fulfill the conditions of Definition 8.26. Then $w_{N,M}$, a solution of the discrete variational inequality* (8.17)*, has the following estimate for $(1-\theta)\delta t(\delta x)^{-2}$ small enough*

$$\max_{0\le n\le N}\|w_M^{(n)}-g_h\|_{L_M^2}^2+C\sum_{n=0}^{N-1}\|w_M^{(n+1)}-w_M^{(n)}\|_{L_M^2}^2+C\sum_{n=0}^{N}\delta t\|w_M^{(n)}\|_{H_M^1}^2 \le CT\|g_h\|_{H_M^1}^2.$$

Proof. Inserting $v_M = g_h$ into (8.17) we have

$$\begin{aligned}&\Big(w_M^{(n+1)}-w_M^{(n)},g_h-w_M^{(n)}\Big)_{L_M^2}\\&-\delta t\Big(\Lambda\big((1-\theta)w_M^{(n+1)}+\theta w_M^{(n)}\big),g_h-w_M^{(n)}\Big)_{L_M^2}\le 0.\end{aligned}\tag{8.18}$$

Using the identities

$$\begin{aligned}&\Big(w_M^{(n+1)}-w_M^{(n)},g_h-w_M^{(n)}\Big)_{L_M^2}\\&\quad=\frac12\Big(\|w_M^{(n)}-g_h\|_{L_M^2}^2+\|w_M^{(n+1)}-w_M^{(n)}\|_{L_M^2}^2-\|w_M^{(n+1)}-g_h\|_{L_M^2}^2\Big),\end{aligned}$$

$$\begin{aligned}\Big(\Lambda\big((1-\theta)w_M^{(n+1)}+\theta w_M^{(n)}\big),w_M^{(n)}\Big)_{L_M^2}&=\theta\Big(\Lambda w_M^{(n)},w_M^{(n)}\Big)_{L_M^2}\\&+(1-\theta)\Big(\Lambda w_M^{(n+1)},w_M^{(n+1)}\Big)_{L_M^2}-(1-\theta)\Big(\Lambda w_M^{(n+1)},w_M^{(n+1)}-w_M^{(n)}\Big)_{L_M^2},\end{aligned}$$

and the estimates of Definition 8.26 we obtain from (8.18)

$$\begin{aligned}&\|w_M^{(n)}-g_h\|_{L_M^2}^2+\|w_M^{(n+1)}-w_M^{(n)}\|_{L_M^2}^2+2\delta t\theta\beta\|w_M^{(n)}\|_{H_M^1}^2\\&\quad+2\delta t(1-\theta)\beta\|w_M^{(n+1)}\|_{H_M^1}^2\\&\quad\le\|w_M^{(n+1)}-g_h\|_{L_M^2}^2+2\delta t(1-\theta)\alpha\|w_M^{(n+1)}\|_{H_M^1}\|w_M^{(n+1)}-w_M^{(n)}\|_{H_M^1}\\&\quad+2\delta t\alpha\Big((1-\theta)\|w_M^{(n+1)}\|_{H_M^1}+\theta\|w_M^{(n)}\|_{H_M^1}\Big)\|g_h\|_{H_M^1}.\end{aligned}$$

Since

$$\|v\|_{H^1_M} \le C_0(\delta x)^{-1}\|v\|_{L^2_M} \tag{8.19}$$

then using the Cauchy inequality we obtain

$$\begin{aligned}\|w_M^{(n)} - g_h\|^2_{L^2_M} &+ \Big(1 - \frac{C_0^2}{2\epsilon}(1-\theta)\alpha\delta t(\delta x)^{-2}\Big)\|w_M^{(n+1)} - w_M^{(n)}\|^2_{L^2_M} \\ &+ 2\delta t\theta(\beta - \alpha\epsilon)\|w_M^{(n)}\|^2_{H^1_M} + 2\delta t(1-\theta)(\beta - 2\alpha\epsilon)\|w_M^{(n+1)}\|^2_{H^1_M} \\ &\le \delta t\frac{\alpha}{2\epsilon}\|g_h\|^2_{H^1_M} + \|w_M^{(n+1)} - g_h\|^2_{L^2_M}.\end{aligned}$$

Taking $\delta t(\delta x)^{-2}$ such small that $\big(1 - \frac{C_0^2}{2\epsilon}(1-\theta)\alpha\delta t(\delta x)^{-2}\big) \ge \gamma > 0$, and ϵ such that $2\alpha\epsilon < \beta$, we obtain

$$\begin{aligned}\|w_M^{(n)} - g_h\|^2_{L^2_M} + \gamma\|w_M^{(n+1)} - w_M^{(n)}\|^2_{L^2_M} + C\delta t\|w_M^{(n)}\|^2_{H^1_M} + C\delta t\|w_M^{(n+1)}\|^2_{H^1_M} \\ \le C\delta t\|g_h\|^2_{H^1_M} + \|w_M^{(n+1)} - g_h\|^2_{L^2_M}.\end{aligned}$$

Summing the above inequality from N down to $m \ge 0$ we obtain

$$\begin{aligned}\|w_M^{(m)} - g_h\|^2_{L^2_M} + \gamma\sum_{n=m}^{N-1}\|w_M^{(n+1)} - w_M^{(n)}\|^2_{L^2_M} \\ + C\sum_{n=m}^{N}\delta t\|w_M^{(n)}\|^2_{H^1_M} + C\sum_{n=m}^{N}\delta t\|w_M^{(n+1)}\|^2_{H^1_M} \\ \le CT\|g_h\|^2_{H^1_M}.\end{aligned}$$

■

In the proof of convergence, we will use the weak formulation of a variational inequality. The theorem below shows that the weak discrete formulation follows from the strong formulation, the result analogous to the corollary of Theorem 8.19.

THEOREM. 8.32 *Let $w_{N,M}$ be a solution of the strong formulation of discrete variational inequality*

$$\begin{aligned}&\sum_{n=0}^{N-1} -\delta t\Big(\frac{w_M^{(n+1)} - w_M^{(n)}}{\delta t}, v_M^{(n)} - w_M^{(n)}\Big)_{L^2_M} \\ &\qquad + \sum_{n=0}^{N-1}\delta t\Big(\Lambda\big((1-\theta)w_M^{(n+1)} + \theta w_M^{(n)}\big), v_M^{(n)} - w_M^{(n)}\Big)_{L^2_M} \ge 0, \\ &w_M^{(n)} \ge g_h,\ n = 0, \dots, N, \\ &w_M^{(N)} = g_h,\end{aligned}$$

for each $v_{N,M}$ such that $v_M^{(n)} \in H_M^1$ and $v_M^{(n)} \geq g_h$, $n = 0, \ldots, N$.

Then $w_{N,M}$ solves the following weak formulation

$$\begin{aligned}
&\sum_{n=0}^{N-1} -\delta t\Big(\frac{v_M^{(n+1)} - v_M^{(n)}}{\delta t}, v_M^{(n)} - w_M^{(n)}\Big)_{L_M^2} \\
&\qquad + \sum_{n=0}^{N-1} \delta t\Big(\Lambda\big((1-\theta)w_M^{(n+1)} + \theta w_M^{(n)}\big), v_M^{(n)} - w_M^{(n)}\Big)_{L_M^2} \\
&\qquad \geq -\frac{1}{2}\|v_M^{(N)} - w_M^{(N)}\|_{L_M^2}, \\
&w_M^{(n)} \geq g_h,\ n = 0, \ldots, N, \\
&w_M^{(N)} = g_h,
\end{aligned}$$

for each $v_{N,M}$ as above.

THEOREM. 8.33 *Let $w_{N,M}$ be a sequence of solutions of the discrete variational inequality* (8.17) *for δt and δx tending to zero ($N, M \to \infty$). Let u be a unique solution of the continuous variational inequality* (8.12) *with the bilinear form $B_R[u, v]$ defined by* (8.13)*. Let $\hat{w}_{N,M}$ denote a piecewise constant extension of $w_{N,M}$. Then for $(1-\theta)\frac{\delta t}{(\delta x)^2} \to 0$ as δt, δx go to zero, we obtain the convergence*

$$\begin{aligned}
&\hat{w}_{N,M} \to u \text{ in } L^2(0, T; L^2(U_R)), \\
&(\delta_x w_{N,M})\hat{} \to Du \text{ in } L^2(0, T; L^2(U_R)).
\end{aligned}$$

Proof. From the stability results of Theorem 8.31 it follows that the sequence $\hat{w}_{N,M}$ is bounded. Then this sequence contains a weakly convergent subsequence, which we also denote $\hat{w}_{N,M}$. Let

$$\begin{aligned}
&\hat{w}_{N,M} \rightharpoonup w \text{ in } L^2(0, T; L^2(U_R)), \\
&(\delta_x w_{N,M})\hat{} \rightharpoonup Dw \text{ in } L^2(0, T; L^2(U_R)).
\end{aligned}$$

Our goal is to prove that w is a solution of the variational inequality (8.12) with $B_R[u, v]$ defined by (8.13). We take $v \in L^2(0, T; H^1(U_R))$ such that $\frac{dv}{dt} \in L^2(0, T; H^{-1}(U_R))$, $v(T) = g$, and $v(t) \geq g$ for $t \in [0, T)$. Let $\hat{v}_{N,M}$ be a sequence of piecewise constant approximations to v which exists due to Lemma

8.27. Inserting $v_{N,M}$ into the weak inequality of Theorem 8.32 we obtain

$$\sum_{n=0}^{N-1} -\delta t\Big(\frac{v_M^{(n+1)}-v_M^{(n)}}{\delta t}, v_M^{(n)}-w_M^{(n)}\Big)_{L_M^2} + \sum_{n=0}^{N-1}\delta t\Big(\Lambda\big((1-\theta)w_M^{(n+1)}+\theta w_M^{(n)}\big), v_M^{(n)}-w_M^{(n)}\Big)_{L_M^2} \geq 0.$$

After a rearrangement we have

$$\sum_{n=0}^{N-1} -\delta t\Big(\frac{v_M^{(n+1)}-v_M^{(n)}}{\delta t}, v_M^{(n)}-w_M^{(n)}\Big)_{L_M^2} + \sum_{n=0}^{N-1}\delta t\Big(\Lambda w_M^{(n)}, v_M^{(n)}-w_M^{(n)}\Big)_{L_M^2} \geq (1-\theta)\sum_{n=0}^{N-1}\delta t\Big(\Lambda\big(w_M^{(n)}-w_M^{(n+1)}\big), v_M^{(n)}-w_M^{(n)}\Big)_{L_M^2}. \tag{8.20}$$

Since $\hat{w}_{N,M}$ converges to w weakly then due to Lemma 8.29 we have

$$\liminf_{\delta t,\delta x\to 0}\int_0^T \big(\Lambda\hat{w}_{N,M},\hat{w}_{N,M}\big)_{L^2(U_R)}dt \geq \int_0^T B_R[w,w]dt.$$

As $\hat{v}_{N,M}\to v$ and $(\delta_t v_{N,M})\hat{}\to \frac{dv}{dt}$ strongly then passing to the limit $\delta t,\delta x\to 0$ on the left hand side of (8.20) (with the change from $v_{N,M}$ and $w_{N,M}$ to $\hat{v}_{N,M}$ and $\hat{w}_{N,M}$) and using the consistency results of Definition 8.28 one gets

$$\int_0^T -\Big\langle\frac{dv}{dt}(t), v(t)-w(t)\Big\rangle dt + \int_0^T B_R\big[w(t),v(t)-w(t)\big]dt \geq \limsup_{\delta t,\delta x\to 0}(1-\theta)\sum_{n=0}^{N-1}\delta t\Big(\Lambda\big(w_M^{(n)}-w_M^{(n+1)}\big), v_M^{(n)}-w_M^{(n)}\Big)_{L_M^2}.$$

It remains to prove that

$$X_{N,M} = (1-\theta)\sum_{n=0}^{N-1}\delta t\Big(\Lambda\big(w_M^{(n)}-w_M^{(n+1)}\big), v_M^{(n)}-w_M^{(n)}\Big)_{L_M^2}$$

tends to zero.

Due to the estimates of Definition 8.26 and (8.19), we obtain

$$\begin{aligned}|X_{N,M}| &\le (1-\theta)\alpha\sum_{n=0}^{N-1}\delta t\|(w_M^{(n)}-w_M^{(n+1)}\|_{H_M^1}\|v_M^{(n)}-w_M^{(n)}\|_{H_M^1}\\ &\le C_0(1-\theta)\alpha(\delta x)^{-1}\sum_{n=0}^{N-1}\delta t\|(w_M^{(n)}-w_M^{(n+1)}\|_{L_M^2}\|v_M^{(n)}-w_M^{(n)}\|_{H_M^1}\\ &\le C_0(1-\theta)\alpha(\delta x)^{-1}(\delta t)^{\frac{1}{2}}\Bigg(\sum_{n=0}^{N-1}\|(w_M^{(n)}-w_M^{(n+1)}\|_{L_M^2}^2\Bigg)^{\frac{1}{2}}\\ &\qquad\times\Bigg(\sum_{n=0}^{N-1}\delta t\|v_M^{(n)}-w_M^{(n)}\|_{H_M^1}^2\Bigg)^{\frac{1}{2}}.\end{aligned}$$

The term

$$\sum_{n=0}^{N-1}\|(w_M^{(n)}-w_M^{(n+1)}\|_{L_M^2}^2$$

is bounded due to Theorem 8.31. Since $v_{N,M}, w_{N,M}\in L_N^2(0,T;H_M^1(U_R))$

$$\sum_{n=0}^{N-1}\delta t\|v_M^{(n)}-w_M^{(n)}\|_{H_M^1}^2$$

is also bounded.

Due to the assumption $(1-\theta)(\delta x)^{-2}\delta t\to 0$ as $\delta t,\delta x\to 0$. Hence $X_{N,M}\to 0$ and we have

$$\int_0^T -\Big\langle\frac{dv}{dt}(t),v(t)-w(t)\Big\rangle dt+\int_0^T B_R\big[w(t),v(t)-w(t)\big]dt\ge 0,$$

which proves that w is a solution of the weak formulation of variational inequality (8.11). Due to Theorem 8.21 a solution of the weak variational inequality (8.11) is also a solution of strong variational inequality (8.12) and since (8.12) possesses a unique solution then $w\equiv u$.

We have to prove that the convergence of $\hat{w}_{N,M}$ to u and $(\delta_x w_{N,M})\hat{}$ to Du is strong in $L^2(0,T;L^2(U_R))$. Let $\hat{u}_{N,M}$ be a piecewise constant approximation of u which exists due to Lemma 8.27. We will show that

$$Y_{N,M}=\sum_{n=0}^{N}\delta t\big(\Lambda w_M^{(n)}-\Lambda u_M^{(n)},w_M^{(n)}-u_M^{(n)}\big)_{L_M^2}$$

converges to zero as $\delta x,\delta t\to 0$.

Taking in the weak variational inequality of Theorem 8.32 $v_{N,M} = u_{N,M}$ we obtain (observe that $u(T) = g$)

$$\begin{aligned}
&\sum_{n=0}^{N-1} \delta t\Big(\Lambda\big((1-\theta)w_M^{(n+1)} + \theta w_M^{(n)}\big), w_M^{(n)}\Big)_{L_M^2} \\
&\qquad \le \sum_{n=0}^{N-1} -\delta t\Big(\frac{u_M^{(n+1)} - u_M^{(n)}}{\delta t}, u_M^{(n)} - w_M^{(n)}\Big)_{L_M^2} \\
&\qquad + \sum_{n=0}^{N-1} \delta t\Big(\Lambda\big((1-\theta)w_M^{(n+1)} + \theta w_M^{(n)}\big), u_M^{(n)}\Big)_{L_M^2}.
\end{aligned} \tag{8.21}$$

Since

$$\begin{aligned}
Y_{N,M} = &\sum_{n=0}^{N} \delta t\big(\Lambda w_M^{(n)}, w_M^{(n)}\big)_{L_M^2} - \sum_{n=0}^{N} \delta t\big(\Lambda w_M^{(n)}, u_M^{(n)}\big)_{L_M^2} \\
&+ \sum_{n=0}^{N} \delta t\big(\Lambda u_M^{(n)}, u_M^{(n)} - w_M^{(n)}\big)_{L_M^2}
\end{aligned} \tag{8.22}$$

and

$$\begin{aligned}
&\sum_{n=0}^{N-1} \delta t\Big(\Lambda\big((1-\theta)w_M^{(n+1)} + \theta w_M^{(n)}\big), w_M^{(n)}\Big)_{L_M^2} \\
&\qquad + (1-\theta) \sum_{n=0}^{N-1} \delta t\Big(\Lambda w_M^{(n)} - \Lambda w_M^{(n+1)}, w_M^{(n)}\Big)_{L_M^2} \\
&\qquad = \sum_{n=0}^{N-1} \delta t\Big(\Lambda w_M^{(n)}, w_M^{(n)}\Big)_{L_M^2}
\end{aligned} \tag{8.23}$$

we can insert (8.21) and (8.23) into (8.22) to obtain

$$\begin{aligned}
Y_{N,M} \le &\sum_{n=0}^{N-1} -\delta t\Big(\frac{u_M^{(n+1)} - u_M^{(n)}}{\delta t}, u_M^{(n)} - w_M^{(n)}\Big)_{L_M^2} \\
&+ \sum_{n=0}^{N-1} \delta t\Big(\Lambda\big((1-\theta)w_M^{(n+1)} + \theta w_M^{(n)}\big), u_M^{(n)}\Big)_{L_M^2} \\
&+ (1-\theta) \sum_{n=0}^{N-1} \delta t\Big(\Lambda w_M^{(n)} - \Lambda w_M^{(n+1)}, w_M^{(n)}\Big)_{L_M^2} \\
&- \sum_{n=0}^{N} \delta t\big(\Lambda w_M^{(n)}, u_M^{(n)}\big)_{L_M^2} + \sum_{n=0}^{N} \delta t\big(\Lambda u_M^{(n)}, u_M^{(n)} - w_M^{(n)}\big)_{L_M^2}.
\end{aligned}$$

Using the equality

$$(1-\theta)w_M^{(n+1)} + \theta w_M^{(n)} = w_M^{(n)} - (1-\theta)\big(w_M^{(n)} - w_M^{(n+1)}\big)$$

we obtain the estimate

$$\begin{aligned} Y_{N,M} \le \sum_{n=0}^{N-1} \delta t\Big(-\frac{u_M^{(n+1)} - u_M^{(n)}}{\delta t} + \Lambda u_M^{(n)}, u_M^{(n)} - w_M^{(n)}\Big)_{L_M^2} \\ - (1-\theta)\sum_{n=0}^{N-1} \delta t\Big(\Lambda w_M^{(n)} - \Lambda w_M^{(n+1)}, u_M^{(n)} - w_M^{(n)}\Big)_{L_M^2}. \end{aligned}$$

We prove now

$$Z = (1-\theta)\sum_{n=0}^{N-1} \delta t\Big(\Lambda w_M^{(n)} - \Lambda w_M^{(n+1)}, u_M^{(n)} - w_M^{(n)}\Big)_{L_M^2} \to 0.$$

From the estimates of Definition 8.26 and (8.19) we have

$$\begin{aligned} |Z| &\le (1-\theta)\alpha \sum_{n=0}^{N-1} \delta t \|w_M^{(n)} - w_M^{(n+1)}\|_{H_M^1} \|u_M^{(n)} - w_M^{(n)}\|_{H_M^1} \\ &\le C_0(1-\theta)\alpha(\delta x)^{-1} \sum_{n=0}^{N-1} \delta t \|w_M^{(n)} - w_M^{(n+1)}\|_{L_M^2} \|u_M^{(n)} - w_M^{(n)}\|_{H_M^1} \\ &\le C(1-\theta)(\delta x)^{-1}(\delta t)^{\frac{1}{2}} \Big(\sum_{n=0}^{N-1} \|w_M^{(n)} - w_M^{(n+1)}\|_{L_M^2}^2\Big)^{\frac{1}{2}} \\ &\qquad \times \Big(\sum_{n=0}^{N-1} \delta t \|u_M^{(n)} - w_M^{(n)}\|_{H_M^1}^2\Big)^{\frac{1}{2}}. \end{aligned}$$

Similarly like above,

$$\sum_{n=0}^{N-1} \|w_M^{(n)} - w_M^{(n+1)}\|_{L_M^2}^2$$

is bounded due to Theorem 8.31, and since $v_{N,M}, w_{N,M} \in L_N^2(0,T;H_M^1(U_R))$ then also

$$\sum_{n=0}^{N-1} \delta t \|u_M^{(n)} - w_M^{(n)}\|_{H_M^1}^2$$

is bounded.

Hence if $(1-\theta)\delta t(\delta x)^{-2} \to 0$ as $\delta t, \delta x \to 0$ then $Z \to 0$.

The remaining expression

$$\sum_{n=0}^{N-1} \delta t\Big(-\frac{u_M^{(n+1)}-u_M^{(n)}}{\delta t}+\Lambda u_M^{(n)}, u_M^{(n)}-w_M^{(n)}\Big)_{L_M^2}$$

also converges to zero. As $\hat{u}_{N,M}\to u$ strongly, hence also weakly, and $\hat{w}_{N,M}\rightharpoonup u$ weakly then $\hat{u}_{N,M}-\hat{w}_{N,M}$ converges to zero weakly. Since $(\delta_t u_{N,M})\hat{}\to \frac{du}{dt}$ strongly then

$$\sum_{n=0}^{N-1} \delta t\Big(-\frac{u_M^{(n+1)}-u_M^{(n)}}{\delta t}, u_M^{(n)}-w_M^{(n)}\Big)_{L_M^2}\to 0.$$

By the strong convergence $\hat{u}_{N,M}\to u$, the weak convergence $\hat{u}_{N,M}-\hat{w}_{N,M}\rightharpoonup 0$, and the consistency result of Lemma 8.29 we have

$$\sum_{n=0}^{N} \delta t\big(\Lambda u_M^{(n)}, u_M^{(n)}-w_M^{(n)}\big)_{L_M^2}=\int_0^T \big(\Lambda\hat{u}_{N,M}, \hat{u}_{N,M}-\hat{w}_{N,M}\big)_{L^2(U_R)}dt\to 0.$$

That proves $Y_{N,M}\to 0$ as $\delta t,\delta x\to 0$.

By the estimates of Definition 8.26 and the convergence $Y_{N,M}\to 0$

$$\sum_{n=0}^{N} \delta t\big(\Lambda w_M^{(n)}-\Lambda u_M^{(n)}, w_M^{(n)}-u_M^{(n)}\big)_{L_M^2}dt\geq \beta\sum_{n=0}^{N}\delta t\|w_M^{(n)}-u_M^{(n)}\|^2_{H_M^1}.$$

Hence

$$\int_0^T \|\hat{w}_{N,M}-\hat{u}_{N,M}\|^2_{L^2(U_R)}dt\to 0,$$

and

$$\int_0^T \|(\delta_x w_{N,M})\hat{}-(\delta_x u_{N,M})\hat{}\|^2_{L^2(U_R)}dt\to 0,$$

which give the desired convergence

$$\hat{w}_{N,M}\to u \text{ strongly in } L^2(0,T;L^2(U_R)),$$
$$(\delta_x w_{N,M})\hat{}\to Du \text{ strongly in } L^2(0,T;L^2(U_R)).$$

■

Projected SOR algorithm

We begin with reformulating the discrete variational inequality (8.17) as a linear complementarity problem. The following lemma is a discrete version of Lemma 8.22.

LEMMA. 8.34 *Let $w_{N,M}$ be a solution of the discrete variational inequality of Definition 8.30. Then $w_{N,M}$ is a solution of the following discrete linear complementarity problem:*
knowing $w_M^{(n+1)} \in H_M^1(U_R)$ such that $w_M^{(n+1)} \geq g_h$, find $w_M^{(n)} \in H_M^1(U_R)$, $n = N-1,\ldots,0$, such that for all $v_M \in H_M^1(U_R)$, $v_M \geq 0$,

$$
\begin{aligned}
&i)\ w_M^{(N)} = g_h,\\
&ii)\ \Big(w_M^{(n+1)} - w_M^{(n)} - \delta t \Lambda\big((1-\theta)w_M^{(n+1)} + \theta w_M^{(n)}\big), v_M\Big)_{L_M^2} \leq 0,\\
&iii)\ w_M^{(n)} \geq g_h,\\
&iv)\ \Big(w_M^{(n+1)} - w_M^{(n)} - \delta t \Lambda\big((1-\theta)w_M^{(n+1)} + \theta w_M^{(n)}\big), g_h - w_M^{(n)}\Big)_{L_M^2} = 0.
\end{aligned}
$$

By the above lemma on each time step $n = N-1,\ldots,0$ we have to solve the linear problem

$$
\begin{aligned}
&QX \geq G,\\
&X \geq \Phi,\\
&(QX - G, X - \Phi) = 0,
\end{aligned}
\tag{8.24}
$$

where $(\cdot,\cdot)$ denotes the scalar product in $\mathbb{R}^{M+1} \equiv L_M^2$ and

$$
\begin{aligned}
Q &= I + \theta\delta t\Lambda,\\
X &= w_M^{(n)},\\
G &= (I - (1-\theta)\delta t\Lambda)w_M^{(n+1)},\\
\Phi &= g_h.
\end{aligned}
$$

Introducing new variables $Z = X - \Phi$ and $V = G - Q\Phi$ we can rewrite the linear complementarity problem in the following form:
find $W = (W_k)_{k=0}^M$ and $Z = (Z_k)_{k=0}^M$ such that

$$
\begin{aligned}
&QZ - W = V,\\
&W \geq 0, \quad Z \geq 0,\\
&(W, Z) = 0.
\end{aligned}
\tag{8.25}
$$

THEOREM. 8.35 *Let Λ fulfill the estimates of Definition 8.26. Then a solution of* (8.25) *is unique.*

Proof. Due to the estimates of Definition 8.26, we have (δt is fixed)

$$C_1\|v\|^2_{H^1_M} \geq (Qv, v)_{L^2_M} \geq C_2\|v\|^2_{H^1_M}.$$

Since now δx is fixed the norms $\|\cdot\|_{L^2_M}$ and $\|\cdot\|_{H^1_M}$ are equivalent. Hence Q is a bounded, positive definite matrix.

Assume for the proof simplicity that Q is symmetric (this assumption will be used further in the PSOR algorithm).

Let us consider the following optimization problem

$$\max_{Z\geq 0} V^\top Z - \frac{1}{2}Z^\top QZ.$$

The Lagrange function for this problem is

$$L(Z, W) = \frac{1}{2}Z^\top QZ - V^\top Z - W^\top Z,$$

where W is a vector of the Lagrange multipliers. Let us observe that the Lagrange function is convex. Hence there is a unique solution of this optimization problem and the solution fulfills the Kuhn-Tucker conditions. Computing the Kuhn-Tucker conditions we obtain

$$\begin{aligned} &QZ - V - W = 0, \\ &W \geq 0, \quad Z \geq 0, \\ &W_k Z_k = 0,\ k = 0, \ldots, M, \text{ hence } (W, Z) = 0. \end{aligned}$$

which is exactly (8.25). ■

We have assumed in the above proof that Q is a symmetric, positive definite matrix. The positive definiteness of matrix Q is essential for the algorithm (see Theorem 8.36). The assumption of symmetry can be partially relaxed to diagonally dominant matrices or a particular type of tridiagonal matrices. We omit such extensions to make the proof of convergence simple.

The PSOR algorithm

Step 1.

Select the relaxation parameter $\omega \in (1, 2)$, the accuracy $\epsilon > 0$, and the starting point $Z^0 \geq 0$. Put $p = 0$.

Step 2.

Compute successively

$$Y_i^{p+1} = V_i - \sum_{j=1}^{i-1} Q_{ij} Z_j^{p+1} - \sum_{j=i}^{M} Q_{ij} Z_j^p,$$
$$Z_i^{p+1} = \max\Big(0, Z_i^p + \omega \frac{Y_i^{p+1}}{Q_{ii}}\Big),$$
$$W_i^{p+1} = -Y_i^{p+1} + Q_{ii}\big(Z_i^{p+1} - Z_i^p\big).$$

Step 3.
If $|Z^{p+1} - Z^p| > \epsilon$ increase p by 1 and return to Step 2, otherwise **stop**. Z^{p+1} is the solution.

THEOREM. 8.36 *Let Z^p, W^p be generated by the PSOR algorithm. Then $Z^p \to Z$, $W^p \to W$ where (Z, W) is a solution of* (8.25).

Proof. Let $F(u) = u^\top Q u - 2u^\top V$. Then by the symmetry of Q

$$F(u) - F(v) = (u - v)^\top Q(u - v) + 2(u - v)^\top (Qv - V).$$

We define the vectors $z^{(p,l)}$, $l = -1, 0, \ldots, M$,

$$z_i^{(p+1,l)} = \begin{cases} Z_i^{p+1}, & \text{for } 0 \le i \le l, \\ Z_i^p, & \text{for } l < i \le M. \end{cases}$$

Then $z^{(p+1,-1)} = Z^p$ and $z^{(p+1,M)} = Z^{p+1}$. With vectors $z^{(p+1,l)}$ we can write

$$Y_i^{p+1} = (V - Q z^{(p+1,i-1)})_i.$$

Let

$$\omega_{(p+1,i)} = \begin{cases} \big(Z_i^{p+1} - Z_i^p\big) \frac{Q_{ii}}{Y_i^{p+1}}, & \text{if } Y_i^{p+1} \neq 0\,, \\ \omega, & \text{if } Y_i^{p+1} = 0. \end{cases}$$

We have $0 \le \omega_{(p+1,i)} \le \omega$ since: $i)$ if $Y_i^{p+1} \neq 0$ and $Z_i^p + \omega \frac{Y_i^{p+1}}{Q_{ii}} \ge 0$ then $\omega_{(p+1,i)} = \omega$; $ii)$ if $Z_i^p + \omega \frac{Y_i^{p+1}}{Q_{ii}} < 0$ then $Y_i^{p+1} < 0$ and $0 \le \omega_{(p+1,i)} < \omega$. Thus we can write

$$Z_i^{p+1} = Z_i^p + \omega_{(p+1,i)} \frac{Y_i^{p+1}}{Q_{ii}}.$$

We have

$$\begin{aligned} F(z^{(p+1,i)}) - F(z^{(p+1,i-1)}) \\ &= \big(z^{(p+1,i)} - z^{(p+1,i-1)}\big)^\top Q\big(z^{(p+1,i)} - z^{(p+1,i-1)}\big) \\ &+ 2\big(z^{(p+1,i)} - z^{(p+1,i-1)}\big)^\top \big(Qz^{(p+1,i-1)} - V\big) \end{aligned}$$

and after simplifications

$$\begin{aligned} F(z^{(p+1,i)}) - F(z^{(p+1,i-1)}) \\ &= Q_{ii}\big(Z_i^{p+1} - Z_i^p\big)^2 - 2\big(Z_i^{p+1} - Z_i^p\big)Y_i^{p+1} \\ &= -\omega_{(p+1,i)}(2 - \omega_{(p+1,i)})\frac{\big(Y_i^{p+1}\big)^2}{Q_{ii}}. \end{aligned}$$

Since $0 \le \omega_{(p+1,i)} \le \omega < 2$ then $F(z^{(p+1,i)}) \le F(z^{(p+1,i-1)})$ and the sequence $F(z^{(p,i)})$ is decreasing. By the definition of $z^{(p,i)}$ we have $F(Z^{p+1}) \le F(Z^p)$. Since $F(u)$ is bounded from below as a quadratic function with matrix Q positive definite then

$$F(Z^p) \searrow F_\infty, \quad p \to \infty.$$

Let $a = \min_{0 \le i \le M} Q_{ii}$ then by the definition of $\omega_{(p+1,i)}$ we obtain

$$\begin{aligned} F(z^{(p+1,i-1)}) - F(z^{(p+1,i)}) &= Q_{ii}\big(Z_i^{p+1} - Z_i^p\big)^2\Big(-1 + \frac{2}{\omega_{(p+1,i)}}\Big) \\ &\ge a\Big(-1 + \frac{2}{\omega}\Big)\big(Z_i^{p+1} - Z_i^p\big)^2. \end{aligned}$$

This gives

$$\big|Z_i^{p+1} - Z_i^p\big| \le \Big(a\Big(-1 + \frac{2}{\omega}\Big)\Big)^{-\frac{1}{2}}\Big(F(z^{(p+1,i-1)}) - F(z^{(p+1,i)})\Big)^{\frac{1}{2}}.$$

The convergence of $F(z^{(p,i)})$ implies

$$\big|Z_i^{p+1} - Z_i^p\big| \to 0, \quad p \to \infty.$$

Let Z be a condensation point of Z^p. There exists a sequence $p_k \to \infty$, for $k \to \infty$, such that

$$Z^{p_k} \to Z, \text{ for } k \to \infty.$$

Then we have

$$\begin{aligned} Y^{p_k} &\to Y = V - QZ, \\ W^{p_k} &\to W = -Y. \end{aligned}$$

We have to prove $Z \geq 0$ and $Y \leq 0$. Condition $Z \geq 0$ follows from the inequality $Z^p \geq 0$ for all p. Let us assume now that $Y > 0$. Then there exist $\epsilon > 0$, p_0, and i_0 such that $Y_{i_0}^{p_k} \geq \epsilon$ for $p_k \geq p_0$. But from Step 2 of the PSOR algorithm we obtain

$$Z_{i_0}^{p_k} - Z_{i_0}^{p_k-1} \geq \frac{\epsilon\omega}{Q_{i_0 i_0}}, \text{ for } p_k \geq p_0$$

and that contradicts the convergence $Z_i^{p_k} - Z_i^{p_k-1} \to 0$.

We will now prove that $Y^\top Z = 0$. Suppose it is not true. Then there exist $\epsilon > 0$, i_0, and p_0 such that

$$Z_{i_0}^{p_k} \geq \epsilon, \quad Y_{i_0}^{p_k} \leq -\epsilon, \text{ for } p_k \geq p_0.$$

Then from Step 2 of the PSOR algorithm

$$Z_{i_0}^{p_k-1} \geq Z_{i_0}^{p_k} \text{ and } \left|Z_{i_0}^{p_k} - Z_{i_0}^{p_k-1}\right| \geq \frac{\epsilon\omega}{Q_{i_0 i_0}}, \text{ for } p_k \geq p_0,$$

which again contradicts the convergence $Z_i^{p_k} - Z_i^{p_k-1} \to 0$.

For the proof completeness, we have to show that the sequence Z^p possesses a condensation point. To this end, let us observe that $Z^p \in R = \{z\colon\ F(z) \leq F(Z^0)\}$ for all p. Since $F(u)$ is bounded from below, R is compact as the inverse image of a compact set under a continuous mapping. Then Z^p has a condensation point as an infinite sequence in a compact set. ■

Penalty method

The basic idea of the penalty method is to replace variational inequality (8.12) by the nonlinear differential equation

$$\begin{aligned} &-\frac{\partial u_\epsilon}{\partial t} + \mathcal{A}^t u_\epsilon + \frac{1}{\epsilon} j(u_\epsilon(t)) = 0, \text{ on } [0,T) \times U_R, \\ &u_\epsilon(t) = g, \text{ on } \partial U_R, \ a.e.\ t \in [0,T), \\ &u_\epsilon(T) = g, \text{ on } U_R, \end{aligned} \tag{8.26}$$

where $j(u) = [u-g]^-$ and $[x]^- = \min(x, 0)$.

We begin the investigation of (8.26) with the following useful lemma.

LEMMA. 8.37 *Operator $j(u)$ is monotone in $L^2(U_R)$, i.e.,*

$$\forall u_1, u_2 \in L^2(U_R) \quad (j(u_1) - j(u_2), u_1 - u_2) \geq 0.$$

This operator is also continuous in $L^2(U_R)$.

Proof. By the elementary estimate

$$(x^- - y^-)(x - y) \geq (x^- - y^-)^2, \quad x, y \in \mathbb{R}$$

we obtain taking $u_1, u_2 \in L^2(U_R)$

$$\begin{aligned}\big(j(u_1) - j(u_2), u_1 - u_2\big) &= \big([u_1 - g]^- - [u_2 - g]^-, u_1 - u_2\big) \\ &= \big([u_1 - g]^- - [u_2 - g]^-, (u_1 - g) - (u_2 - g)\big) \\ &\geq \|[u_1 - g]^- - [u_2 - g]^-\|^2 = \|j(u_1) - j(u_2)\|^2 \geq 0.\end{aligned}$$

The same estimate proves that $j(u)$ is continuous. ∎

From Chapter 3 of the book by Bensoussan and Lions [5] we have the following two theorems.

THEOREM. 8.38 *Let Assumption 8.16 hold in U_R. There exists a unique weak solution $u_\epsilon(t)$ of* (8.26) *with $u_\epsilon \in L^2(0,T;H^1(U_R))$ and $\frac{du_\epsilon}{dt} \in L^2(0,T;L^2(U_R))$.*

THEOREM. 8.39 *Let $u(t)$ be a solution of* (8.12) *with $u \in L^2(0,T;H^1(U_R))$ and $\frac{du}{dt} \in L^2(0,T;L^2(U_R))$. Let $u_\epsilon(t)$ solve* (8.26) *with $u_\epsilon \in L^2(0,T;H^1(U_R))$ and $\frac{du_\epsilon}{dt} \in L^2(0,T;L^2(U_R))$. Then*

$$\max_{t\in[0,T]} \|u(t) - u_\epsilon(t)\|_{L^2(U_R)} \leq C\sqrt{\epsilon},$$

for $C > 0$.

We will now construct a numerical solution of (8.26). Similarly, like for variational inequalities, we limit considerations to a one-dimensional case. We assume that operator $\mathcal{A}^t$ is time-independent and given by formula (8.14) and consider the weak formulation of the differential equation (8.26)

$$\big(-\tfrac{d}{dt}u_\epsilon(t), v\big)_{L^2(U_R)} + B_R\big[u_\epsilon(t), v\big] + \frac{1}{\epsilon}\big(j(u_\epsilon(t)), v\big)_{L^2(U_R)} = 0, \qquad (8.27)$$

which holds $t \in [0,T)$ a.e. for each $v \in H^1(U_R)$.

We approximate the differential problem by finite differences. Like previously, we define space and time grids, the corresponding functional spaces $L^2_M(U_R)$, $H^1_M(U_R)$ and $L^2_N(0,T;L^2_M(U_R))$, $L^2_N(0,T;H^1_M(U_R))$, and grid functions v_M, $v_{N,M}$. Matrix Λ is given by Definition 8.26, and due to Lemma 8.29 $(\Lambda v_M, w_M)$ is consistent with $B_R[v,w]$.

We approximate problem (8.26) for a space-time grid function $w_{N,M}$ by the following θ-scheme

$$w_M^{(n)} - w_M^{(n+1)} + \delta t\Lambda\big((1-\theta)w_M^{(n+1)} + \theta w_M^{(n)}\big) + \frac{\delta t}{\epsilon} j_h(w_M^{(n)}) = 0, \qquad (8.28)$$

where $j_h(w_M^{(n)}) = [w_M^{(n)} - g_h]^-$.

It can be shown under additional mild assumptions (cf. Theorem 8.44) that (8.28) possesses a unique solution.

THEOREM. 8.40 *Let matrix Λ fulfill the conditions of Definition 8.26. Then $w_{N,M}$, a solution of the discrete penalty problem* (8.28)*, has for $(1-\theta)\delta t(\delta x)^{-2}$ small enough the following estimate*

$$\max_{0\le n\le N} \|w_M^{(n)} - g_h\|_{L_M^2}^2 + C\sum_{n=0}^{N-1} \|w_M^{(n+1)} - w_M^{(n)}\|_{L_M^2}^2 + C\sum_{n=0}^{N} \delta t \|w_M^{(n)}\|_{H_M^1}^2 \le CT\|g_h\|_{H_M^1}^2.$$

Proof. Multiplying equation (8.28) scalarly by $(w_M^{(n)} - g_h)$ we obtain

$$\Big(w_M^{(n)} - w_M^{(n+1)}, w_M^{(n)} - g_h\Big)_{L_M^2} + \delta t\Big(\Lambda\big((1-\theta)w_M^{(n+1)} + \theta w_M^{(n)}\big), w_M^{(n)} - g_h\Big)_{L_M^2} + \frac{\delta t}{\epsilon}\Big(j_h(w_M^{(n)}), w_M^{(n)} - g_h\Big)_{L_M^2} = 0. \tag{8.29}$$

We have

$$\Big(j_h(w_M^{(n)}), w_M^{(n)} - g_h\Big)_{L_M^2} = \Big(j_h(w_M^{(n)}) - j_h(g_h), w_M^{(n)} - g_h\Big)_{L_M^2} \ge 0$$

by the monotonicity of j_h which follows from Lemma 8.37 and the observation that $j_h(g_h) = 0$.

Then (8.29) can be reduced to

$$\Big(w_M^{(n)} - w_M^{(n+1)}, w_M^{(n)} - g_h\Big)_{L_M^2} + \delta t\Big(\Lambda\big((1-\theta)w_M^{(n+1)} + \theta w_M^{(n)}\big), w_M^{(n)} - g_h\Big)_{L_M^2} \le 0$$

and this is exactly inequality (8.18) from the proof of Theorem 8.31. Thus the rest of the proof goes exactly like the proof of Theorem 8.31. ∎

With the above estimate we can prove that solutions of problem (8.28) converge to a weak solution of problem (8.26) as $\delta t, \delta x \to 0$.

THEOREM. 8.41 *Let $w_{N,M}$ be a sequence of solutions of the discrete penalty problem* (8.28) *for δt and δx tending to zero ($N, M \to \infty$). Let u_ϵ be a unique*

weak solution of the continuous penalty problem (8.26)*. Let* $\hat{w}_{N,M}$ *denote a piecewise constant extension of* $w_{N,M}$*. Then for* $(1-\theta)\frac{\delta t}{(\delta x)^2}$ *sufficiently small to fulfill the conditions of Theorem 8.40, we obtain the convergence*

$$\hat{w}_{N,M} \rightharpoonup u_\epsilon \text{ in } L^2(0,T;L^2(U_R)),$$
$$(\delta_x w_{N,M})\hat{} \rightharpoonup Du_\epsilon \text{ in } L^2(0,T;L^2(U_R)).$$

Proof. From the stability results of Theorem 8.40 it follows that the sequence $\hat{w}_{N,M}$ is bounded. Then this sequence contains a weakly convergent subsequence, which we also denote $\hat{w}_{N,M}$. Let

$$\hat{w}_{N,M} \rightharpoonup w \text{ in } L^2(0,T;L^2(U_R)),$$
$$(\delta_x w_{N,M})\hat{} \rightharpoonup Dw \text{ in } L^2(0,T;L^2(U_R)).$$

Our goal is to prove that w is a weak solution of the penalty problem (8.26). This proof is analogous to the proof of Theorem 5.28. We take $v \in L^2(0,T;H^1(U_R))$ such that $v(t) \geq g$ for $t \in [0,T]$. Let $\hat{v}_{N,M}$ be a sequence of piecewise constant approximations of v. We multiply (8.28) scalarly by $\hat{v}_M^{(n)}$ and sum over n

$$\sum_{n=0}^{N-1}\Big(\Big(-\frac{\hat{w}_M^{(n+1)}-\hat{w}_M^{(n)}}{\delta t},\hat{v}_M^{(n)}\Big)_{L_M^2}+\Big(\Lambda\big((1-\theta)\hat{w}_M^{(n+1)}+\theta\hat{w}_M^{(n)}\big),\hat{v}_M^{(n)}\Big)_{L_M^2} + \frac{1}{\epsilon}\big(j_h(\hat{w}_M^{(n)}),\hat{v}_M^{(n)}\big)_{L_M^2}\Big)=0.$$

Passing to the limit $\delta t, \delta x \to 0$, using the consistency of $(\Lambda w_M, v_M)$ with $B_R[w,v]$ and the continuity (hence also the weak continuity) of $j(x)$ we obtain

$$\int_0^T\Big(\big(-\tfrac{d}{dt}w(t),v(t)\big)_{L^2(U_R)}+B_R[w(t),v(t)]+\frac{1}{\epsilon}\big(j(w(t)),v(t)\big)_{L^2(U_R)}\Big)dt=0.$$

Since $v \in L^2(0,T;H^1(U_R))$ is arbitrary, we have

$$\big(-\tfrac{d}{dt}w(t),z\big)_{L^2(U_R)}+B_R[w(t),z]+\frac{1}{\epsilon}\big(j(w(t)),z\big)_{L^2(U_R)}=0,$$

for each $z \in H^1(U_R)$. Hence $w(t)$ is a weak solution of (8.26) with $w \geq g$ as all $\hat{w}_{N,M} \geq g_h$ and $w(T) = g$ since $w_M^{(N)} = g_h$. ∎

Remark. 8.6 *Delicate analytical considerations exploring the regularity of solutions of penalty problem* (8.26) *and the estimates of Theorem 8.40 reveal that the weakly convergent sequence* $\hat{w}_{N,M}$ *of Theorem 8.41 is in fact converging strongly to* $w = u_\epsilon$ *in* $L^2(0,T;L^2(U_R))$.

DEFINITION. 8.42 *A square matrix A is called the M-matrix if non-diagonal entries of A are nonpositive ($a_{ij} \leq 0$, $i \neq j$) and all principal minors of A are positive definite. This property can be expressed in terms of entries of A saying that $\forall i\ a_{ii} \geq 0$, $a_{ij} \leq 0$, $i \neq j$ and $\forall i\ \sum_j a_{ij} \geq 0$ with at least one i_0 such that $\sum_j a_{i_0 j} > 0$.*

The property of an M-matrix A which will be particularly important for our future considerations is the positivity of A^{-1}.

Remark. 8.7 *For the Black-Scholes model we have matrix Λ with $a^2 = \frac{1}{2}\sigma^2$, $b = \frac{1}{2}\sigma^2 - r$ and $c = r$. Then Λ is a tridiagonal matrix with*

$$\begin{aligned}
\Lambda_{k,k} &= \sigma^2 \frac{1}{(\delta x)^2} + r, \\
\Lambda_{k,k+1} &= -\frac{1}{2}\sigma^2 \frac{1}{(\delta x)^2} + \Big(\frac{1}{2}\sigma^2 - r\Big)\frac{1}{\delta x}, \\
\Lambda_{k,k-1} &= -\frac{1}{2}\sigma^2 \frac{1}{(\delta x)^2} - \Big(\frac{1}{2}\sigma^2 - r\Big)\frac{1}{\delta x}.
\end{aligned}$$

Hence Λ is an M-matrix for $\delta x < \min(1, \frac{\sigma^2}{2r})$.

Obviously also $Q = (I + \delta t\theta\Lambda)$ is an M-matrix. Q is an M-matrix even for negative interest rates provided δt is sufficiently small.

Denoting

$$\begin{aligned}
Q &= I + \theta\delta t\Lambda, \\
X &= w_M^{(n)}, \\
Q_p &= (I - (1-\theta)\delta t\Lambda), \\
X_p &= w_M^{(n+1)}, \\
\Phi &= g_h,
\end{aligned}$$

we can write (8.28) in the compact form

$$QX + \frac{\delta t}{\epsilon}[X - \Phi]^- = Q_p X_p. \tag{8.30}$$

We will solve (8.30) assuming that Q is an M-matrix. To solve equation (8.30), we will apply an iterative method. But first, we will prove the following estimate.

LEMMA. 8.43 *Let $w_{N,M}$ be a solution of* (8.28)*. Then for $0 \leq n \leq N$*

$$\|[w_M^{(n)} - \Phi]^-\|_{L_M^2} \leq C\frac{\epsilon}{\delta t}.$$

Proof. We denote $Y_\epsilon = w_M^{(n)} - \Phi$. Then (8.30) can be written as

$$QY_\epsilon + \frac{\delta t}{\epsilon}[Y_\epsilon]^- = Q_p X_p - Q\Phi. \tag{8.31}$$

Without loss of generality we can assume that $Y_\epsilon = (y_1, y_2)$, where $y_1 = [Y_\epsilon]^-$, and $y_2 = [Y_\epsilon]^+$, and matrix Q is decomposed accordingly

$$Q = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix}.$$

Let $Q_p X_p - Q\Phi = b$. Due to Theorem 8.40 $\|b\|_{L_M^2}$ is bounded for all $w_{N,M}$.

We multiply scalarly (8.31) by $[Y_\epsilon]^-$ to obtain

$$(Q_{11}y_1, y_1)_{L_M^2} + (Q_{12}y_2, y_1)_{L_M^2} + \frac{\delta t}{\epsilon}\|[Y_\epsilon]^-\|_{L_M^2}^2 = (b, [Y_\epsilon]^-)_{L_M^2}. \tag{8.32}$$

$(Q_{11}y_1, y_1)_{L_M^2} \geq 0$ since Q is an M-matrix. We have $y_1 \leq 0$ and $Q_{12}y_2 \leq 0$ since $y_2 \geq 0$ and $Q_{12} \leq 0$. Then $(Q_{12}y_2, y_1)_{L_M^2} \geq 0$. Dropping these nonnegative terms from the left hand side of (8.32) we obtain

$$\|[Y_\epsilon]^-\|_{L_M^2}^2 \leq \frac{\epsilon}{\delta t}\big|(b, [Y_\epsilon]^-)_{L_M^2}\big| \leq \frac{\epsilon}{\delta t}\|[Y_\epsilon]^-\|_{L_M^2}\|b\|_{L_M^2}.$$

■

A solution of (8.30) is obtained by nonlinear iterations. To define these iterations let us introduce the diagonal matrix

$$P(X)_{ij} = \begin{cases} 1 & \text{if } X < \Phi \text{ and } i = j, \\ 0 & \text{otherwise.} \end{cases} \tag{8.33}$$

Then we can write equation (8.30) as

$$\Big(Q + \frac{\delta t}{\epsilon}P(X)\Big)X = Q_p X_p + \frac{\delta t}{\epsilon}P(X)\Phi.$$

The above equation can be solved by the following Newton iterations:

$$\begin{aligned} &X^0 = X_p, \\ &\Big(Q + \frac{\delta t}{\epsilon}P(X^k)\Big)X^{k+1} = Q_p X_p + \frac{\delta t}{\epsilon}P(X^k)\Phi. \end{aligned} \tag{8.34}$$

The iteration process stops when for a given tolerance η

$$\frac{\|X^{k+1} - X^k\|_{L^\infty(U_R)}}{\max\big(1, \|X^{k+1}\|_{L^\infty(U_R)}\big)} < \eta, \text{ or } P(X^{k+1}) = P(X^k).$$

THEOREM. 8.44 *Let Λ be an M-matrix. Then*

a) the iterates (8.34) *converge to a unique solution of* (8.30) *for any initial value X_p;*

b) the iterates converge monotonically: if $X^1 \geq X^0$ then $X^{k+1} \geq X^k$, for $k \geq 1$.

Proof. *Monotone convergence.* Subtracting from equation (8.34) for X^{k+1} the same equation for X^k we obtain

$$\Big(Q + \frac{\delta t}{\epsilon} P(X^k)\Big)(X^{k+1} - X^k) = \frac{\delta t}{\epsilon}\big(P(X^k) - P(X^{k-1})\big)(\Phi - X^k).$$

Let us examine the term $\big(P(X^k) - P(X^{k-1})\big)(\Phi - X^k)$:

1. if $X_i^k < \Phi_i$ then $P(X^k)_{ii} = 1$ and $\big(P(X^k) - P(X^{k-1})\big)_{ii}(\Phi - X^k)_i \geq 0$;
2. if $X_i^k \geq \Phi_i$ then $P(X^k)_{ii} = 0$ and $\big(P(X^k) - P(X^{k-1})\big)_{ii}(\Phi - X^k)_i = -P(X^{k-1})_{ii}(\Phi - X^k)_i \geq 0$.

Thus we always have

$$\big(P(X^k) - P(X^{k-1})\big)(\Phi - X^k) \geq 0, \quad k \geq 1.$$

Since Λ is an M-matrix then also $(I + \theta\delta t\Lambda + \frac{\delta t}{\epsilon} P(X^k))$ is an M-matrix. Thus for $k \geq 1$ we have

$$(X^{k+1} - X^k) = \Big(I + \theta\delta t\Lambda + \frac{\delta t}{\epsilon} P(X^k)\Big)^{-1}\big(P(X^k) - P(X^{k-1})\big)(\Phi - X^k) \geq 0.$$

Bounded iterates. From (8.34) we have for $k \geq 1$

$$\|X^{k+1}\|_{L^\infty(U_R)} \leq \Big\|\Big(Q + \frac{\delta t}{\epsilon} P(X^k)\Big)^{-1}\Big\|_{\mathcal{L}(L^\infty, L^\infty)} \Big\|Q_p X_p + \frac{\delta t}{\epsilon} P(X^k)\Phi\Big\|_{L^\infty(U_R)} \leq C,$$

since matrix $\Big(Q + \frac{\delta t}{\epsilon} P(X^k)\Big)^{-1}$ is bounded as the inverse of an M-matrix and matrices Q_p and $P(X^k)$ are bounded (in the norm of any finite dimensional space).

The bounded, monotone sequence of iterates converges to a solution of (8.30).

Uniqueness. Assume that there are two solutions of (8.30): V^1 and V^2. Subtracting equation (8.30) for V^2 from the same equation for V^1 we obtain

$$\Big(Q + \frac{\delta t}{\epsilon} P(V^2)\Big)(V^1 - V^2) = \frac{\delta t}{\epsilon}\big(P(V^1) - P(V^2)\big)(\Phi - V^1).$$

Using similar computations as in the proof of monotonicity, we have

$$\big(P(V^1) - P(V^2)\big)(\Phi - V^1) \geq 0.$$

Since $\big(Q + \frac{\delta t}{\epsilon} P(V^2)\big)$ is an M-matrix then $(V^1 - V^2) \geq 0$. Performing the same computations with the role of V^1 and V^2 interchanged we get $(V^2 - V^1) \geq 0$. Hence $V^1 = V^2$. ■

Bibliography

[1] Y. Achdou and O. Pironneau – *Computational Methods for Option Pricing*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2005.

[2] L. Andersen and J. Andreasen – Volatility skews and extensions of the Libor Market Model, *Appl. Math. Finance*, **7** (2000), 1–32.

[3] S. Asmussen and P. W. Glynn – *Stochastic Simulation: Algorithms and Analysis*, Springer, New York, 2007.

[4] J. D. Beasley and S. G. Springer – The percentage points of the normal distribution, *Appl. Stat. J. Roy. St. C*, **26** (1977), 118–121.

[5] A. Bensoussan and J.L. Lions – *Applications of Variational Inequalities in Stochastic Control*, North-Holland, Amsterdam, 1982.

[6] J. F. Bonnans – *Numerical Analysis of Partial Differential Equations Arising in Finance*, Lecture notes, available on the web page of the author, 2018.

[7] B. Bouchard – *Méthodes de Monte Carlo en Finance*, Lecture notes, available on the web page of the author, 2007.

[8] P. Boyle – Options: a Monte Carlo approach, *J. Financ. Econ.*, **4** (1977), 323–338.

[9] H. Brézis – Inéquations variationnelles paraboliques, Séminaire Jean Leray (1971), exp. No. 7, p. 1–10.

[10] M. Broadie, and J. Detemple – American option valuation: new bounds, approximations, and a comparison of existing methods *Rev. Financ. Stud.*, **9** (1996), 1211–1250.

[11] E. Clément, D. Lamberton, and P. Protter – An analysis of the Longstaff-Schwartz algorithm for American option pricing, *Financ. Stoch.*, **6** (2002), 449–471.

[12] R. Courant, K. Friedrichs, and H. Lewy – Über die partiellen Differenzengleichungen der mathematischen Physik, *Math. Ann.*, **100**(1), (1928), 32–74.

[13] J. C. Cox, S. A. Ross, and M. Rubinstein – Option pricing: A simplified approach, *J. Financ. Econ.*, **7** (1979), 229–263.

[14] C.W. Cryer – The solution of a quadratic programming problem using systematic overrelaxation, *SIAM J. Control*, **9** (1971), 385–392.

[15] S. Dekel and D. Leviatan – The Bramble–Hilbert lemma for convex domains, *SIAM J. Math. Anal.*, **35**(5) (2004), 1203–1212.

[16] J. Dick and F. Pillichshammer – *Digital Nets and Sequences*, Cambridge University Press, 2010.

[17] D. Egloff – Monte Carlo algorithms for optimal stopping and statistical learning, *Ann. Appl. Probab.*, **15** (2005), 1396–1432.

[18] N. El Karoui – Les aspects probabilistes du contrôle stochastique, in *Lecture Notes in Mathematics* **876**, Springer, New York, 1981, pp. 72–238.

[19] L. C. Evans – *Partial Differential Equations*, American Mathematical Society, Providence, RI, 1998.

[20] P. A. Forsyth and K. R. Vetzal – Quadratic convergence for valuing American options using a penalty method, *SIAM J. Sci. Comput.*, **23** (2002), 2095–2122.

[21] P. Glasserman – *Monte Carlo Methods in Financial Engineering*, Springer, New York, 2004.

[22] R. Glowinski, J.-L. Lions and R. Trémollières – *Analyse numérique des inéquations variationnelles*, Dunod, Paris, 1976.

[23] D. J. Higham – Nine ways to implement the binomial method for option valuation in MATLAB, *SIAM Rev.*, **44** (2002), 661–677.

[24] N. Hilber, O. Reichmann, C. Schwab and C. Winter – *Computational Methods for Quantitative Finance*, Springer, Berlin, 2013.

[25] P. Jäckel – *Monte-Carlo Methods in Finance*, Wiley Finance Series, John Wiley & Sons, Chichester, 2002.

[26] P. Jaillet, D. Lamberton, and B. Lapeyre – Variational inequalities and the pricing of American options, *Acta Appl. Math.*, **21** (1990), 263–289.

[27] R. Jarrow and A. Rudd – *Option Pricing*, Richard D. Irwin, Homewood, Ill., 1983.

[28] B. Kamrad and P. Ritchken – Multinomial approximating models for options with k-state variables, *Manage. Sci.*, **37**(12) (1991), 1640–1652.

[29] I. Karatzas and S. E. Shreve – *Brownian Motion and Stochastic Calculus*, Springer, New York, second edition, 1991.

[30] P. E. Kloeden and E. Platen – *Numerical Solution of Stochastic Differential Equations*, Springer, New York, 1992.

[31] R. Korn, E. Korn and G. Kroisandt – *Monte Carlo Methods and Models in Finance and Insurance*, CRC Press, Boca Raton, 2010.

[32] N. V. Krylov – *Introduction to the Theory of Diffusion Processes*, American Mathematical Society, Providence, RI, 1995.

[33] H. Kunita – *Stochastic Differential Equations and Stochastic Flows of Diffeomorphisms* in *Lecture Notes in Mathematics* **1097**, Springer, Berlin, 1984, pp. 143–303.

[34] P. L'Ecuyer – Efficient and portable combined random number generators, *Commun. ACM*, **31** (1988), 742–749.

[35] S. Larsson and V. Thomée – *Partial Differential Equations with Numerical Methods*, Springer, Berlin, 2009.

[36] J. London – *Modeling Derivatives in C++*, Wiley Finance Series, John Wiley & Sons, New Jersey, 2005.

[37] F. A. Longstaff and E. S. Schwartz – Valuing American options by simulation: a simple least-squares approach, *Rev. Financ. Stud.*, **14** (2001), 113–147.

[38] G. Marsaglia – Random numbers fall mainly in the planes, *Proc. Nat. Acad. Sci. USA*, **61** (1968), 23–28.

[39] M. Matsumoto and T. Nishimura – Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator, *ACM Trans. Model. Comput. Simul.*, **8**(1) (1998), 3–30.

[40] G. N. Milstein and M. V. Tretyakov – *Stochastic Numerics for Mathematical Physics*, Springer, Berlin, 2004.

[41] B. Moro – The full monte, *Risk*, **8**(2) (1995), 57–58.

[42] K. W. Morton and D. F. Mayers – *Numerical Solution of Partial Differential Equations*, Cambridge University Press, 2005.

[43] H. R. Neave – On using the Box-Muller transformation with multiplicative congruential pseudo-random number generators, *Appl. Stat. J. Roy. St. C*, **22** (1973), 92–97.

[44] A. Quarteroni – *Numerical Models for Differential Problems*, Springer Italia, second edition, 2014.

[45] A. Quarteroni and A. Valli – *Numerical Approximation of Partial Differential Equations*, Springer, Berlin, 1994.

[46] M. Schroder – Computing the constant elasticity of variance option pricing formula, *J. Finance*, **44**(1) (1989), 211–219.

[47] R. Seydel – *Tools for Computational Finance*, Sixth Edition, Springer, London, 2017.

[48] W. T. Shaw, T. Luu and N. Brickman – Quantile mechanics II: Changes of variables in Monte Carlo methods and GPU-optimized normal quantiles, http://arxiv.org/abs/0901.0638v5 (ver. October 23, 2018).

[49] L. Stentoft – Convergence of the least squares Monte Carlo approach to American option valuation, *Manage. Sci.*, **50**(9) (2004), 1193–1203.

[50] J. C. Strikwerda – *Finite Difference Schemes and Partial Differential Equations*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2004.

[51] J. W. Thomas –*Numerical Partial Differential Equations: Finite Difference Methods*, Springer, Berlin, 1995.

[52] V. Thomée – *Galerkin Finite Element Methods for Parabolic Problems*, Springer, Berlin, second edition, 2006.

[53] D. Zanger – Convergence of a least-squares Monte Carlo algorithm for bounded approximating sets, *Appl. Math. Finance*, **16** (2009), 123–150.

[54] D. Zanger – Quantitative error estimates for a least-squares Monte Carlo algorithm for American option pricing, *Financ. Stoch.*, **17** (2013), 503–534.

[55] X. L. Zhang – Numerical analysis of American option pricing in a jump-diffusion model, *Math. Oper. Res.*, **22** (1997), 668–690.