Stochastic Differential Equations: Models and Numerics [1]

Jesper Carlsson Kyoung-Sook Moon Anders Szepessy
Raúl Tempone Georgios Zouraris

February 12, 2019

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

Contents

Chapter 1

Introduction to Mathematical Models and their Analysis

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

Stochastic and deterministic differential equations are fundamental for the modeling in Science and Engineering. As the computational power increases, it becomes feasible to use more accurate differential equation models and solve more demanding problems: for instance to determine input data from fundamental principles, to optimally reconstruct input data using measurements or to find the optimal construction of a design. There are therefore two interesting computational sides of differential equations:

- the forward problem, to accurately determine solutions of differential equations for given data with minimal computational work and prescribed accuracy, and
- the inverse problem, to determine the input data for differential equations, from optimal estimates, based either on measurements or on computations with a more fundamental model.

The model can be stochastic by different reasons:

- if calibration of data implies this, as in financial mathematics, or
- if fundamental microscopic laws generate stochastic behavior when coarse-grained, as in molecular dynamics for chemistry, material science and biology, or
- if a deterministic problem is more efficiently computed by a stochastic method, as for stochastic gradient descent in machine learning or Monte Carlo sampling of high dimensional integrals.

An understanding of which model and method should be used in a particular situation requires some knowledge of both the model approximation error and the discretization error of the method. The optimal method clearly minimizes the computational work for given accuracy. Therefore it is valuable to know something about computational accuracy and work for different numerical models and methods, which lead us to error estimates and convergence results. In particular, our study will take into account the amount of computational work for alternative mathematical models and numerical methods to solve a problem with a given accuracy.

1.1 Noisy Evolution of Stock Values

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

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

which has the solution

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

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

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

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

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

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

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

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

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

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

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

$$f_t \approx \frac{f(t_{n+1}) - f(t_n)}{\Delta t},$$

while the Monte Carlo method discretizes a probability space by substituting expected values with averages of finite samples, e.g. $\{S(T, \omega_j)\}_{j=1}^{M}$ and

$$f(s,t) \approx \sum_{j=1}^{M} \frac{e^{-r(T-t)} \max(S(T, \omega_j) - K, 0)}{M}.$$

Which method is best? The solution depends on the problem to solve and we will carefully study qualitative properties of the numerical methods to understand the answer.

1.2 Molecular Dynamics

An example where the noise can be derived from fundamental principles is molecular dynamics, modeling e.g. reactions in chemistry and biology. Theoretically molecular systems can be modeled by the Schrödinger equation

$$i\partial_t \Psi = H\Psi$$

where the unknown Ψ is a wave function depending on time t and the variables of coordinates and spins of all, M, nuclei and, N, electrons in the problem; and H is the Hamiltonian precisely defined by well known fundamental constants of nature and the Coulomb interaction of all nuclei and electrons. An important issue is its high computational complexity for problems with more than a few nuclei, due to the high dimension of Ψ which is roughly in $L^2(\mathbb{R}^{3(M+N)})$, see [26]. Already simulation of a single water molecule requires a partial differential equation in 39 space dimensions, which is a demanding task to solve also with modern sparse approximation techniques.

A substantial dimensional reduction is obtained with Born-Oppenheimer approximation treating the nuclei as classical particles with the electrons in the ground state

corresponding to the current nuclei positions. This approximation, derived from a WKB approximation for heavy nuclei mass (see Section **??**), leads to *ab initio* molecular dynamics

$$\begin{aligned} \dot{x}_t =& v_t, \\ m\dot{v}_t =& -V'(x_t). \end{aligned} \tag{1.5}$$

To determine the nuclei dynamics and find the electron energy (input to V) means now to solve a differential equation in $\mathbb{R}^{6M}$ where at each time step the electron ground state energy needs to be determined for the current nuclei configuration x^t, see [26, 19]. To simulate large systems with many particles requires some simplification of the expensive force calculation $\partial_{x_i} V$ involving the current position $x_t \in \mathbb{R}^{3M}$ of all nuclei.

The Hamiltonian system (1.5) is often further modified. For instance, equation (1.5) corresponds to simulate a problem with the number of particles, volume and total energy held constant. Simulation of a system with constant number of particles, volume and temperature are often done by using (1.5) and regularly rescaling the kinetic energy to meet the fixed temperature constraint, using so called thermostats. A mathematically attractive alternative to approximate a system in constant temperature is to solve the Langevin-Itô stochastic differential equation

$$\begin{aligned} dx_t &= v_t dt, \\ m dv_t &= -(V'(x_t) + \tau^{-1} v_t)dt + (2k_B T \tau^{-1})^{1/2} dW_t \end{aligned} \tag{1.6}$$

where T is the temperature, k_B the Boltzmann constant, W is a standard Wiener process in $\mathbb{R}^{3M}$ and τ is a relaxation time parameter (which can be determined from molecular dynamics simulation). The Langevin model (1.6) can be derived from the Schrödinger equation under certain assumptions, which is the subject of Sections **??** to **??**. If diffusion is important in the problem under study, one would like to make long simulations on times of order at least τ^{-1}. A useful observation to efficiently simulate longer time is the fact that for $\tau \to 0+$ the solution $x_{s/\tau}$ of the Langevin equation (1.6) converges to the solution $\bar{x}_s$ solving the Smoluchowski equation, also called Brownian dynamics

$$d\bar{x}_s = -V'(\bar{x}_s)ds + (2k_B T)^{1/2} d\bar{W}_s, \tag{1.7}$$

set in the slower diffusion time scale $s = \tau t$. Here, for simplicity, the mass is assumed to be the same for all particles and normalized to $m = 1$ and $\bar{W}$ is again a standard Wiener process in $\mathbb{R}^{3M}$. The Smoluchowski model hence has the advantage to be able to approximate particle systems over longer time and reducing to half the problem dimension by eliminating the velocity variables. In Section **??** we analyze the weak approximation error $x_{s/\tau} \rightharpoonup \bar{x}_s$. The next step in the coarse-graining process is to derive partial differential equations – for the mass, momentum and energy of a continuum fluid – from Langevin or Smoluchowski molecular dynamics, which determines the otherwise unspecified pressure, viscosity and heat conductivity; Section **??** shows an example of such a coarse-graining process in the case of modelling a solid-liquid melt.

1.3 Optimal Control of Investments

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

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

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

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

where $H(u, u_x, u_{xx}) := \max_\alpha \left((\mu\alpha x + r(1 - \alpha)x)u_x + \sigma^2\alpha^2x^2u_{xx}/2\right)$. Part of our work is to study the theory of Hamilton Jacobi equations and numerical methods for control problems to determine the Hamiltonian H and the control α. It turns out that typically the Hamiltonian needs to slightly modified in order to compute an approximate solution: Chapter 8 explains why and how. We call such modifications regularizations.

Chapter 8 also includes a study on rare events for stochastic differential equations, e.g. the important problem of determining reaction rates and reaction path in molecular dynamics, with small noise term. The analysis of these rare events in Chapter 8 are also based on Hamilton Jacobi equations.

1.4 Calibration of the Volatility

Another important application of optimal control we will study is to solve inverse problems for differential equations in order to determine the input data for the differential equation from observed solution values, such as finding the volatility in the Black-Scholes equation from observed option prices: the option values can be used to detemine the volatility function implicitly. The objective in the optimal control formulation is then to find a volatility function that yields option prices that deviate as little as possible from the measured option prices. The dynamics is the Black-Scholes equation with the volatility function to be determined, that is the dynamics is a determinstic partial differential equation and the volatility is the control function, see Chapter 8.2.1.1. This is a typical inverse problem: it is called inverse because in the standard view of the Black-Scholes equation relating the option values and the volaility, the option price is the unknown and the volatility is the data; while here the formulation is reversed with option prices as data and volatility as unknown in the same Black-Scholes equation. Inverse problems are often harder to solve than the forward problem and need to regularized as explained in Chapter 8.

1.5 The Coarse-graining and Discretization Analysis

Our analysis of models and discretization methods use only one basic idea, which we present here for a determinstic problem of two differential equations

$$\dot{X}^t = a(X^t)$$

and

$$\dot{\bar{X}}^t = \bar{a}(\bar{X}^t).$$

We may think of the two given fluxes a and $\bar{a}$ as either two different differential equation models or two discretization methods. The goal is to estimate a quantity of interest $g(X^T)$, e.g. the potential energy of a molecular dynamic system, the lift of an airfoil or the contract of a contingent claim in financial mathematics. Consider therefore a given function $g : \mathbb{R}^d \to \mathbb{R}^d$ with a solution $X : [0,T] \to \mathbb{R}^d$, e.g. the coordinates of atoms in a molecular system or a discretization of mass, momentum and energy of a fluid. To understand the global error $g(X^T) - g(\bar{X}^T)$ we introduce the value function $\bar{u}(x,t) := g(\bar{X}^T;\ \bar{X}^t = x)$, which solves the partial differential equation

$$\begin{aligned} \partial_t \bar{u}(x,t) + \bar{a}(x)\partial_x \bar{u}(x,t) = 0 \quad t < T \\ u(\cdot, T) = g \end{aligned} \tag{1.8}$$

This definition and telescoping cancellation imply that the global error has the representation

$$\begin{aligned} g(X^T) - g(\bar{X}^T) &= \bar{u}(X^T, T) - \bar{u}(\underbrace{\bar{X}^0}_{=X^0}, 0) \\ &= \bar{u}(X^T, T) - \bar{u}(X^0, 0) \\ &= \int_0^T d\bar{u}(X^t, t) \\ &= \int_0^T \partial_t \bar{u}(X^t, t) + \dot{X}^t \partial_x \bar{u}(X^t, t)\, dt \\ &= \int_0^T \partial_t \bar{u}(X^t, t) + \bar{a}(X^t, t)\partial_x \bar{u}(X^t, t)\, dt \\ &= \int_0^T \big(-\bar{a}(X^t, t) + a(X^t, t)\big)\partial_x \bar{u}(X^t, t)\, dt. \end{aligned} \tag{1.9}$$

Here we can identify the local error in terms of the residual $-\bar{a}(X^t,t)+\bar{a}(X^t,t)$ multiplied by the weight $\partial_x \bar{u}(X^t,t)$ and summed over all time steps. Note that the difference of the two solutions in the global error is converted into a weighted average of the residual $-\bar{a}(X^t,t) + \bar{a}(X^t,t)$ along only one solution X^t; the representation is therefore the residual of X-path inserted into the $\bar{u}$-equation. We may view the error representation as a weak form of Lax Equivalence result, which states that the combination of consistence and stability imply convergence: consistence means that the flux $\bar{a}$ approximates a; stability means that $\partial_x \bar{u}$ is bounded in some sense; and convergence means that the global error $g(X^T) - g(\bar{X}^T)$ tends to zero. The equivalence, as it is usually known, is

stated using bounds with appropriate norms and it has been the basis of the theoretical understanding of numerical methods.

The weak formulation (1.9) is easy to use and it is our basis for understanding both modelling and discretization errors. The weak form is particularly useful for estimating the weak approximation error, since it can take cancellation into account by considering the weaker concept of the value function instead of using absolute values and norms of differences of solution paths; the standard strong error analysis is obtained by estimating the norm of the difference of the two paths X and $\bar{X}$. Another attractive property of the weak representation (1.9) is that it can be applied both in *a priori* form to give qualitative results, by combining it with analytical estimates of $\partial_x \bar{u}$, and in *a posteriori* form to obtain also quantitative results, by combining it with computer based estimates of $\partial_x \bar{u}$.

We first use the representation for understanding the weak approximation of stochastic differential equations and its time discretization, by extending the chain rule to Ito's formula and integrate over all outcomes (i.e. take the expected value). The value function solves a parabolic diffusion equation in this case, instead of the hyperbolic transport equation (1.8).

In the case of coarse-graining and modelling error, the representation is used for approximating

- Schrödinger dynamics by stochastic molecular Langevin dynamics,
- Kinetic Monte Carlo jump dynamics by SDE dynamics,
- Langevin dynamics by Smoluchowski dynamics, and
- Smoluchowski molecular dynamics by continuum phase-field dynamics.

We also use the representation for the important problem to analyse inverse problems, such as callibrating the volatility for stocks by observed option prices or finding an optimal portfolio of stocks and bonds. In an optimal control setting the extension is then to include a control parameter α in the flux so that

$$\dot{X}^t = a(X^t, \alpha^t)$$

where the objective now is to find the minimum $\min_\alpha g(X^T;\ X^t = x) =: u(x,t)$. Then the value function u solves a nonlinear Hamilton-Jacobi-Bellman equation and the representation is extended by including a minimum over α.

1.6 Machine Learning

Here is first a short description of a machine learning problem to determine a neural network function from given data. For example, we are given data $\{(x_n, y_n)\}_{n=1}^N$, where $(x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ are independent samples drawn from an unknown probability

density on $\mathbb{R}^d \times \mathbb{R}$. The objective is to train/learn a neural network function, e.g. $\alpha(x,\theta) := \sum_{k=1}^K \theta_k^1 \sigma(\theta_k^2 \cdot x + \theta_k^3)$, that solves the minimization problem

$$\min_{\theta \in \mathbb{R}^{(d+2)K}} \mathbb{E}[f\big(\alpha(x,\theta), y\big)]$$

with the activation function $\sigma(y) := 1/(1+e^{-y})$, the loss function $f(\alpha, y) := |\alpha - y|^2$ and the neural network parameters $\theta = (\theta_k^1, \theta_k^2, \theta_k^3)_{k=1}^K$ where $\theta_k^1 \in \mathbb{R}, \theta_k^2 \in \mathbb{R}^d$ and $\theta_k^3 \in \mathbb{R}$. The stochastic gradient descent method for the iterations $\theta[n] \in \mathbb{R}^{(d+2)K}$, $n = 0, 1, 2, \ldots$ satisfying

$$\begin{aligned} \theta[0] &= \text{ some random guess in } \mathbb{R}^{(d+2)K}, \\ \theta[n+1] &= \theta[n] - \Delta t \nabla_\theta f\big(\alpha(x_n, \theta[n]), y_n\big), \ n = 0, 1, 2, \ldots \end{aligned} \tag{1.10}$$

is often used to approximately solve this minimization problem, based on a step size/learning rate $\Delta t > 0$. By writing the stochastic gradient descent method as

$$\begin{aligned} \theta[0] &= \text{ some random guess in } \mathbb{R}^{(d+2)K}, \\ \theta[n+1] &= \theta[n] - \Delta t \nabla_\theta \mathbb{E}[f\big(\alpha(x_n, \theta[n]), y_n\big)] \\ &\quad + \Delta t \Big(\mathbb{E}[\nabla_\theta f\big(\alpha(x_n, \theta[n]), y_n\big)] - \nabla_\theta f\big(\alpha(x_n, \theta[n]), y_n\big) \Big), \ n = 0, 1, 2, \ldots \end{aligned}$$

it can be understood as a Euler approximation of a certain stochastic differential equation with drift $\nabla_\theta \mathbb{E}[f\big(\alpha(x_n, \theta[n]), y_n\big)]$ and a small noise term. We take advantage of this correspondence between stochastic differential equations and the stochastic gradient descent method in the study on machine learning in Chapter 10. The convergence towards the minimum involves approximation related to time steps, as in (1.2), Monte Carlo sampling, and time asymptotic convergence towards the equilibrium density and the rare events using Hamilton Jacobi equations. The theory of numerical approximation of stochastic differential equations presented here is therefore particularly suited for basic mathematical understanding for machine learning algorithms, as we shall see in Chapter 10.

Chapter 2

Stochastic Integrals

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

2.1 Probability Background

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

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

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

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

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

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

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

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

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

Definition 2.3. A random variable X, in the probability space $(\Omega, \mathcal{F}, P)$, is a function $X : \Omega \to \mathbb{R}^d$ such that the inverse image

$$X^{-1}(A) \equiv \{\omega \in \Omega : X(\omega) \in A\} \in \mathcal{F},$$

for all open subsets A of $\mathbb{R}^d$.

Definition 2.4 (Independence of random variables)**.** Two sets $A, B \in \mathcal{F}$ are said to be independent if

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

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

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

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

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

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

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

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

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

which also can be written

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

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

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

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

2.2 Brownian Motion

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

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

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

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

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

are *independent*; and

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

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

Does there exists a Wiener process and how to construct W if it does? In computations we will only need to determine W at finitely many time steps $\{t_n : \ n = 0, \ldots, N\}$ of the form $0 = t_0 < t_1 < \ldots < t_N = T$. The definition then shows how to generate $W(t_n)$ by a sum of independent normal distributed random variables, see Example 2.20 for computational methods to generate independent normal distributed random variables. These independent increments will be used with the notation $\Delta W_n = W(t_{n+1}) - W(t_n)$. Observe, by Properties 1 and 3, that for fixed time t the Brownian motion $W(t)$ is itself a normal distributed random variable. To generate W for all $t \in \mathbb{R}$ is computationally infeasible, since it seems to require infinite computational work. Example 2.20 shows the existence of W by proving uniform convergence of successive continuous piecewise linear approximations. The approximations are based on an expansion in the orthogonal $L^2(0, T)$ Haar-wavelet basis.

2.3 Approximation and Definition of Stochastic Integrals

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

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

Taking expected values we obtain by Property 2 of Definition 2.8

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

Now let us use instead the *backward Euler* discretization

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

Taking expected values yields a different result:

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

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

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

□

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

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

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

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

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

with the corresponding forward Euler approximations

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

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

Let the maximum time step Δt_{max} *be*

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

Then

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

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

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

Then in that grid we can write

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

where $\Delta f_k = f(\bar{t}_n, W(\bar{t}_n)) - f(\bar{\bar{t}}_m, W(\bar{\bar{t}}_m))$, $\Delta W_k = W(t_{k+1}) - W(t_k)$ and the indices m, n satisfy $t_k \in [\bar{\bar{t}}_m, \bar{\bar{t}}_{m+1})$ and $t_k \in [\bar{t}_n, \bar{t}_{n+1})$, as depicted in Figure 2.1.

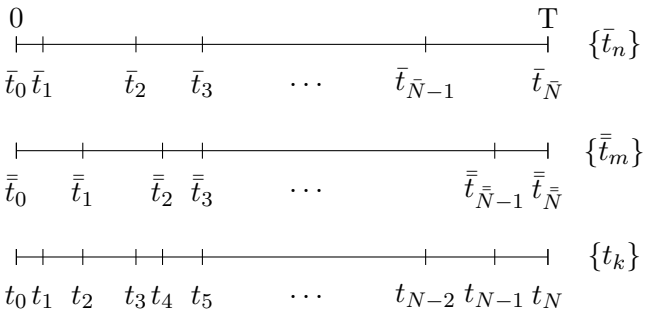


Figure 2.1: Mesh points used in the proof.

Therefore,

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

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

$(a+b)^2 \leq 2(a^2+b^2)$. Substituting this in (2.7) proves the theorem

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

□

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

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

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

Exercise 2.12. Use the forward Euler discretization to show that

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

Example 2.13 (Discrete Wiener process)**.** A discrete Wiener process can be simulated by the following Octave/Matlab code:

```
% Simulation of Wiener process/Brownian path

N = 1E6;                        % number of timesteps
randn('state',0);               % initialize random number generator
T = 1;                          % final time
dt = T/(N-1);                   % time step
t = 0:dt:T;
dW = sqrt(dt)*randn(1,N-1);     % Wiener increments
W = [0 cumsum(dW)];             % Brownian path
```

Brownian paths resulting from different seeds is shown in Figure 2.2, and in e.g. Exercise 2.12, the integrals can then be evaluated by

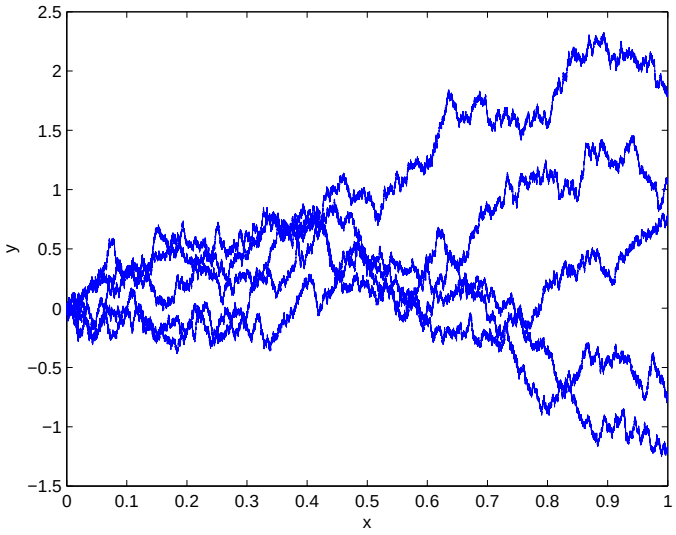


Figure 2.2: Brownian paths

```
LHS = sum(t(1:N-1).*dW);
RHS = T*W(N) - sum(W(1:N-1))*dt;
```

Definition 2.14. A process $f : [0,T] \times \Omega \to \mathbb{R}$ is *adapted* if $f(t,\cdot)$ only depends on events which are generated by $W(s)$, $s \leq t$.

Remark 2.15 (Extension to adapted Itô integration)**.** Itô integrals can be extended to adapted processes. Assume $f : [0,T] \times \Omega \to \mathbb{R}$ is adapted and that there is a constant C such that

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

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

Theorem 2.16 (Basic properties of Itô integrals)**.**
Suppose that $f, g : [0,T] \times \Omega \to \mathbb{R}$ are Itô integrable, e.g. adapted and satifying (2.9), and that c_1, c_2 are constants in $\mathbb{R}$. Then:

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

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

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

Proof. To verify Property (ii), we first use that f is adapted and the independence of the increments ΔW_n to show that for an Euler discretization

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

It remains to verify that the limit of Euler discretizations preserves this property: Cauchy's inequality and the convergence result (2.6) imply that

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

Property (i) and (iii) can be verified analogously. □

Example 2.17 (The Monte-Carlo method)**.** To verify Property (ii) in Theorem 2.16 numerically for some function f we can do a Monte-Carlo simulation where

$$\int_0^T f(s,\cdot)dW(s),$$

is calculated for several paths, or *realizations*, and then averaged:

```
% Monte-Carlo simulation

N = 1E3;                        % number of timesteps
randn('state',0);               % initialize random number generator
T = 1;                          % final time
dt = T/N;                       % time step
t = 0:dt:T;
M = 1E6;                        % number of realisations
MC = zeros(1,M);                % vector to hold mean values

for i=1:M
  dW = sqrt(dt)*randn(1,N);          % Wiener increments
  W = [0 cumsum(dW)];                % Brownian paths
  f = t.^3.*sqrt(abs(W));            % some function
  int = sum(f(1:N).*dW);             % integral value
  if i==1
    MC(i) = int;
  else
    MC(i) = (MC(i-1)*(i-1)+int)/i; % new mean value
  end
end
```

In the above code the mean value of the integral is calculated for $1, \dots, M$ realizations, and in Figure 2.3 we see that as the number of realizations grows, the mean value approaches zero as $1/\sqrt{M}$. Also, from the proof of Theorem 2.16 it can be seen that the number of time steps does not affect this convergence, so the provided code is inefficient, but merely serves as an illustration for the general case.

Exercise 2.18. Use the forward Euler discretization to show that

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

(b) Property (i) and (iii) in Theorem 2.16 hold.

Exercise 2.19. Consider the Ornstein-Uhlenbeck process defined by

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

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

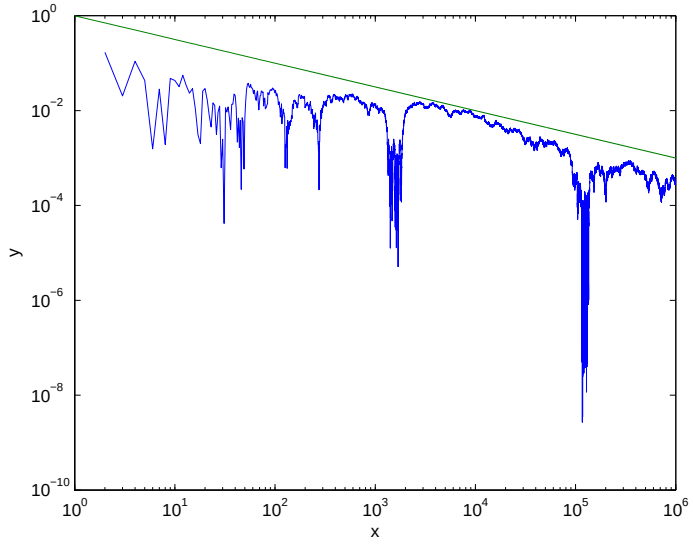


Figure 2.3: Absolute value of the mean for different number of realizations.

Example 2.20 (Existence of a Wiener process)**.** To construct a Wiener process on the time interval $[0,T]$, define the Haar-functions H_i by $H_0(t) \equiv 1$ and for $2^n \leq i < 2^{n+1}$ and $n = 0, 1, 2 \ldots$, by

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

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

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

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

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

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

The functions S_i are small "hat"-functions with a maximum value $T^{-1/2}2^{-(n+2)/2}$ and zero outside an interval of length $T2^{-n}$. Let us postpone the proof that $W^{(m)}$ converge uniformly and first assume this. Then the limit $W(t) = \sum_{i=1}^{\infty} \xi_i S_i(t)$ is continuous. To

verify that the limit W is a Wiener process, we first observe that $W(t)$ is a sum of normal distributed variables so that $W(t)$ is also normal distributed. It remains to verify that the increments ΔW_n and ΔW_m are independent, for $n \neq m$, and $E[(\Delta W_n)^2] = \Delta t_n$. Parseval's equality shows the independence and the correct variance

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

To prove uniform convergence, the goal is to establish

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

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

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

If

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

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

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

but

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

so that $\sum_{n=0}^{\infty} P(\chi_n > n) < \infty$, which contradicts (2.14) and (2.13). Therefore

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

which proves the uniform convergence. □

Exercise 2.21 (Extension to multidimensional Itô integrals)**.** The multidimensional Wiener process W in $\mathbb{R}^l$ is defined by $W(t) \equiv (W^1(t), \dots, W^l(t))$, where $W^i,\ i = 1, \dots, l$ are independent one-dimensional Wiener processes. Show that

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

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

Exercise 2.22. Generalize Theorem 2.16 to multidimensional Itô integrals.

Remark 2.23. A larger class of Itô integrable functions are the functions in the Hilbert space

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

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

Exercise 2.24. Read Example 2.20 and show that the Haar-functions can be used to approximate stochastic integrals $\int_0^T f(t) dW(t) \simeq \sum_{i=0}^m \xi_i f_i$, for given deterministic functions f with $f_i = \int_0^T f(s) H_i(s) ds$. In what sense does $dW(s) = \sum_{i=0}^\infty \xi_i H_i ds$ hold?

Exercise 2.25. Give an interpretation of the approximation (2.12) in terms of Brownian bridges, cf. [24].

Chapter 3

Stochastic Differential Equations

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

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

3.1 Approximation and Definition of SDE

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

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

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

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

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

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

$$d\bar{X}(t) = \bar{a}(t, \bar{X})dt + \bar{b}(t, \bar{X})dW(t), \ \bar{t}_n \leq t < \bar{t}_{n+1}, \quad (3.3)$$

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

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

$$dX(t) = a(t, X(t))dt + b(t, X(t))dW(t), \ 0 \leq t < T, \quad (3.4)$$

with time steps

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

respectively, and

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

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

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

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

and

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

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

Then there is a constant K such that

$$\max\left\{E[\bar{X}^2(t,\cdot)], E[\bar{\bar{X}}^2(t,\cdot)]\right\} \le K(T+1), \ t < T, \tag{3.9}$$

and

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

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

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

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

Then

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

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

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

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

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

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

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

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

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

(3) Use the property (3) from Theorem 2.16.

(4) Apply Grönwall's lemma.

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

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

The definition of the discretized solutions implies that

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

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

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

Note that (3.7) and (3.8) imply

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

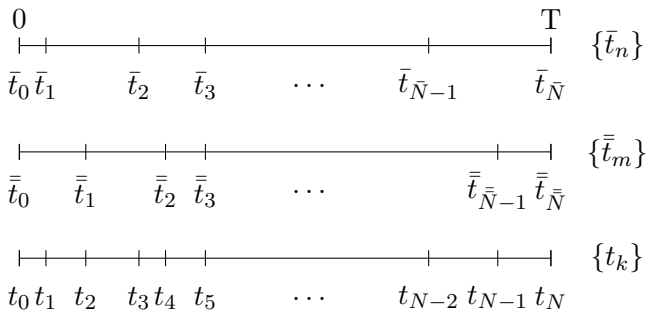


Figure 3.1: Mesh points used in the proof.

Therefore

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

The combination of (3.13) and (3.15) shows

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

and in a similar way,

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

and by the assumptions (3.7)

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

Therefore, the last three inequalities imply

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

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

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

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

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

Similarly, we have

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

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

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

Observe that both the functions $\Delta a(t)$ and $\Delta b(t)$ are adapted and piecewise constant on the refined grid. The error representation (3.12) and (3) of Theorem 2.16 imply

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

Inequalities (3.17-3.19) combine to

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

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

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

which finishes the proof. □

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

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

Remark 3.5 (Strong and weak convergence)**.** Depending on the application, our interest will be focused either on strong convergence

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

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

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

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

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

Corollary 3.4 shows that successive refinements of the forward Euler approximation forms a Cauchy sequence in the Hilbert space V, defined by Definition 2.23. The limit $X \in V$, of this Cauchy sequence, satisfies the stochastic equation

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

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

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

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

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

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

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

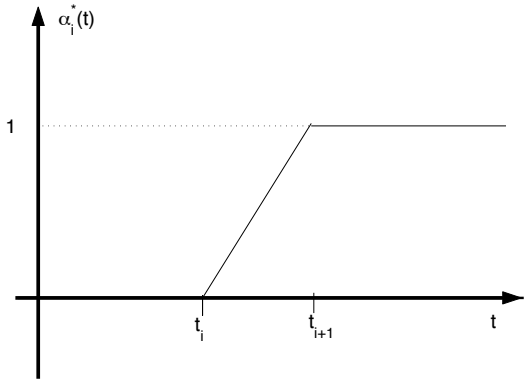


Figure 3.2: Optimal choice for weight functions α_i.

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

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

where $\alpha_i : [0,1] \to \mathbb{R}$, $i = 1,\dots,N$ are any functions. The idea is to choose the functions $\alpha_i : [0,1] \to \mathbb{R}$, $i = 1,\dots,N$ in an optimal way, and see that the minimum error satisfies (3.23). We have

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

and in addition

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

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

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

and hence

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

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

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

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

3.2 Itô's Formula

Recall that using a forward Euler discretization we found the relation

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

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

Theorem 3.9. *Suppose that the assumptions in Theorem 3.1 hold and that X satisfies the stochastic differential equation*

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

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

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

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

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

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

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

Taylor expansion of g up to second order gives

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

The combination of (3.27) and (3.28) shows

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

Let us first show that

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

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

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

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

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

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

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

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

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

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

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

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

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

and

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

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

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

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

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

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

Exercise 3.15. Consider the stochastic differential equation

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

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

(i) Using that

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

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

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

(iii) Use an integrating factor to derive the exact solution (2.10) in Example 2.19. Compare your results from (i) and (ii) with this exact solution.

Example 3.16. Consider the stochastic differential equation

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

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

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

so that

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

and consequently

$$S(T) = e^{(r-\frac{\sigma^2}{2})T + \sigma W(T)} S(0). \tag{3.31}$$

Example 3.17 (Verification of strong and weak convergence)**.** From the explicit formula (3.31) we can numerically verify the results on strong and weak convergence, given in Remark 3.5 for the Euler method. In the following code we calculate the strong and weak error by comparing the Euler simulation and the explicit value (3.31) at final time for several realizations. This is then tested for different time steps and the result in Figure 3.3 confirms a strong convergence of order 1/2 and a weak convergence of order 1.

```
% Stong and weak convergence for the Euler method

steps = [1:6];
  for i=steps
    N = 2^i                          % number of timesteps
    randn('state',0);
    T = 1; dt = T/N; t = 0:dt:T;
    r = 0.1; sigma = 0.5; S0 = 100;
    M = 1E6;                         % number of realisations
    S = S0*ones(M,1);                % S(0) for all realizations
    W = zeros(M,1);                  % W(0) for all realizations
    for j=1:N
      dW = sqrt(dt)*randn(M,1);   % Wiener increments
      S = S + S.*(r*dt+sigma*dW); % processes at next time step
      W = W + dW;                 % Brownian paths at next step
    end
  ST = S0*exp( (r-sigma^2/2)*T + sigma*W ); % exact final value
  wError(i) = mean(S-ST));                  % weak error
  sError(i) = sqrt(mean((S-ST).^2));        % strong error
end
dt = T./2^steps;
loglog(dt,abs(wError),'o--',dt,dt,'--',dt,abs(sError),'o-',dt,sqrt(dt))
```

Exercise 3.18. Suppose that we want to simulate $S(t)$, defined in the previous example by means of the forward Euler method, i.e.

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

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

$$P(S_{n+1} < 0 | S_n = s) < \epsilon \ll 1. \tag{3.32}$$

Motivate a choice for ϵ and find then the largest Δt_n satisfying (3.32).

Remark 3.19. The Wiener process has unbounded variation i.e.

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

This is the reason why the forward and backward Euler methods give different results.

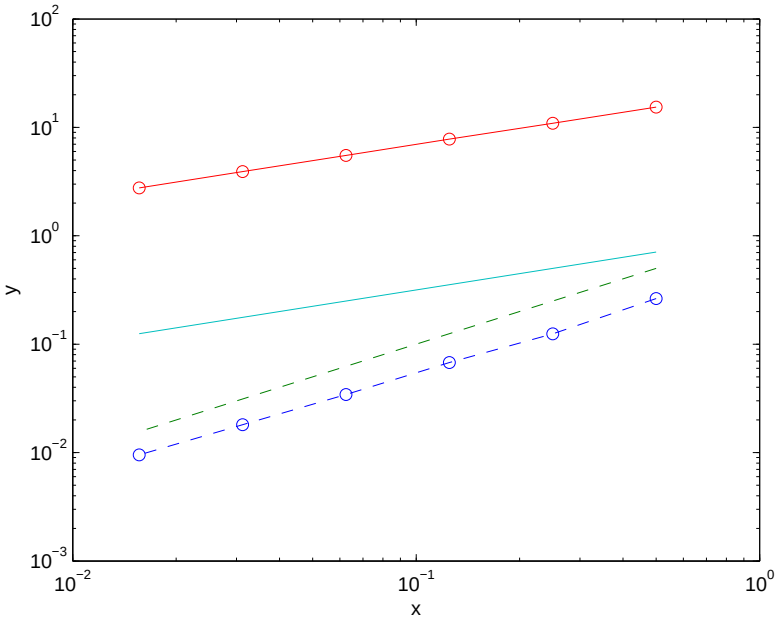


Figure 3.3: Strong and weak convergence.

We have for a uniform mesh $\Delta t = T/N$

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

3.3 Stratonovich Integrals

Recall from Chapter 2 that Itô integrals are constructed via forward Euler discretizations and Stratonovich integrals via the trapezoidal method, see Exercise 3.20. Our goal here is to express a Stratonovich integral

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

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

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

Then the relation reads

$$\begin{aligned}\int_0^T g(t,X(t)) \circ dW(t) &= \int_0^T g(t,X(t))dW(t) \\ &+ \frac{1}{2}\int_0^T \partial_x g(t,X(t))b(t,X(t))dt. \end{aligned} \tag{3.33}$$

Therefore, Stratonovich integrals satisfy

$$dg(t,X(t)) = \partial_t g(t,X(t))dt + \partial_x g(t,X(t)) \circ dX(t), \tag{3.34}$$

just like in the usual calculus.

Exercise 3.20. Use that Stratonovich integrals $g(t,X(t)) \circ dW(t)$ are defined by limits of the trapezoidal method to verify (3.33), cf. Remark 2.9.

Exercise 3.21. Verify the relation (3.34), and use this to show that $dS(t) = rS(t)dt + \sigma S(t) \circ dW(t)$ implies $S(t) = e^{rt+\sigma W(t)}S(0)$.

Remark 3.22 (Stratonovich as limit of piecewise linear interpolations)**.** Let $R^N(t) \equiv W(t_n) + \frac{W(t_{n+1})-W(t_n)}{t_{n+1}-t_n}(t-t_n)$, $t \in (t_n, t_{n+1})$ be a piecewise linear interpolation of W on a given grid, and define X^N by $dX^N(t) = a(X^N(t))dt + b(X^N(t))dR^N(t)$. Then $X^N \to X$ in L^2, where X is the solution of the Stratonovich stochastic differential equation

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

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

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

so that

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

The limit $N \to \infty$ implies $X^N(t) \to X(t) = e^{rt+\sigma W(t)}X(0)$, as in Exercise 3.21.

3.4 Systems of SDE

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

$$dX_i(t) = a_i(t,X(t))dt + b_{ij}(t,X(t))dW_j(t), \ \ for\ i = 1 \ldots d. \tag{3.35}$$

Here and below we use of the summation convention

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

i.e., if the same summation index appears twice in a term, the term denotes the sum over the range of this index. Theorem 3.9 can be directly generalized to the system (3.35).

Theorem 3.23 (Itô 's formula for systems)**.** *Let*

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

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

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

or in matrix vector notation

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

Remark 3.24. The formal rules to remember Theorem 3.23 are Taylor expansion to second order and

$$\begin{aligned} dW_j dt \ &= \ dt dt = 0 \\ dW_i dW_j \ &= \ \delta_{ij} dt = \begin{cases} dt & if\ i = j, \\ 0 & otherwise. \end{cases} \end{aligned} \tag{3.36}$$

Exercise 3.25. Verify Remark 3.24.

Chapter 4

The Feynman-Kăc Formula and the Black-Scholes Equation

4.1 The Feynman-Kăc Formula

Theorem 4.1. *Suppose that a, b and g are differentiable to any order and these derivatives are bounded. Let X be the solution of the stochastic differential equation,*

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

and let $u(x,t) = E[g(X(T))|X(t) = x]$. Then u is the solution of the Kolmogorov backward equation

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

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

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

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

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

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

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

Therefore

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

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

Exercise 4.2 (Maximum Principle)**.** Let the function u satisfy

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

Prove that u satisfies the maximum principle

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

4.2 Black-Scholes Equation

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

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

$$\begin{aligned} dI &= -df \;+\; \alpha dS \;+\; \beta dB \\ &= -(f_t \;+\; \mu Sf_S \;+\; \frac{1}{2}\sigma^2S^2f_{SS})dt \;-\; f_S\sigma SdW+\alpha(\mu Sdt+\sigma SdW)+\beta rBdt \\ &= \left(-(f_t \;+\; \mu Sf_S \;+\; \frac{1}{2}\sigma^2S^2f_{SS}) \;+\; (\alpha\mu S \;+\; \beta rB)\right)dt \;+\; (-f_S \;+\; \alpha)\sigma SdW. \end{aligned}$$

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

$$\begin{aligned} dI &= \left(-(f_t \;+\; \mu Sf_S \;+\; \frac{1}{2}\sigma^2S^2f_{SS}) \;+\; (f_S\mu S \;+\; \beta rB)\right)dt \\ &= \left(-(f_t \;+\; \frac{1}{2}\sigma^2S^2f_{SS}) \;+\; \beta rB\right)dt. \end{aligned} \tag{4.2}$$

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

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

Equation (4.2) and (4.3) show that

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Then u is the solution of the partial differential equation

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

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

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

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

$$\begin{aligned} E[g(X(t))G(T) \ \ | \ \ X(t) = x] - \hat{u}(x,t) & \\ = \ E[\int_t^T G L^* \hat{u} \ ds] \ + \ E[\int_t^T b G \hat{u}_x \ dW] \ + \ E[\int_t^T \hat{u} V G \ ds] & \\ = \ E[\int_t^T h G \ ds] \ - \ E[\int_t^T \hat{u} V G \ ds] \ + \ E[\int_t^T \hat{u} V G \ ds] & \\ = \ E[\int_t^T h G \ ds | X(t) = x]. & \end{aligned}$$

Therefore

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

□

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

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

$$\underbrace{-\partial_s P(y,s;x,t) - \partial_y \big(a(y,s) P(y,s;x,t)\big) + \frac{1}{2} \partial_y^2 \big(b^2(y,s) P(y,s;x,t)\big)}_{=: LP} = 0, \ \ s > t$$

$$P(y,t;x,t) = \delta(x-y),$$

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

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

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

Consequently,

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

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

Exercise 4.10 (Limit probability distribution)**.** Consider the Ornstein-Uhlenbeck process defined by

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

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

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

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

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

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

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

Chapter 5

The Monte-Carlo Method

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

5.1 Statistical Error

Consider the stochastic differential equation

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

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

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

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

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

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

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

Solution. We have

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

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

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

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

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

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

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

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

and

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

The representation (5.3) implies

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

Therefore

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

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

Exercise 5.3. Show that (5.4) implies

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

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

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

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

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

Solution. Let the error ϵ_N be defined by

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

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

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

In practice, σ^2 is approximated by

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

□

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

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

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

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

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

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

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

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

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

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

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

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

and

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

Example 5.7. Consider the system of stochastic differential equations,

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

Then

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

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

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

5.2 Time Discretization Error

Consider the stochastic differential equation

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

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

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

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

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

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

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

Theorem 5.8. *Assume that a, b and g are differentiable to any order and these derivatives are bounded, then there holds*

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

Proof. Let u satisfy the equation

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

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

The assumptions can be used to verify that u and its derivatives exists. We do not verify this here. The Feynman-Kăc formula shows

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

and in particular

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

Then by the Itô formula,

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

Evaluate the integral from 0 to T,

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

Take the expected value and use (5.10) to obtain

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

The following Lemma 10.5 proves the Theorem. □

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

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

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

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

Provided $|f_1'(t)| \le C$, the initial condition (10.45) implies that $f_1(t) = \mathcal{O}(\Delta t_n)$, for $t_n \le t < t_{n+1}$. Therefore, it remains to show that $|f_1'(t)| \le C$. Let $\alpha(t,x) = -(a(t,x) - a(t_n, \overline{X}(t_n)))u_x(t,x)$, so that $f(t) = E[\alpha(t, \overline{X}(t))]$. Then by Itô 's formula

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

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

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

Similarly, we can also prove

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

□

Example 5.10. Consider the stochastic volatility model,

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

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

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

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

$$dS = rSdt + \sigma SdW(t),$$

with constant interest rate r and volatility σ.

(i) Show that

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

(ii) Use equation (5.13) to simulate the price

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

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

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

(iv) Estimate the accuracy of your results. Suggest a better method to solve this problem.

Exercise 5.12. Assume that a system of stocks solves

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

where W_j are independent Brownian motions.

(i) Show that

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

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

Exercise 5.13 (An example of variance reduction)**.** Consider the computation of a call option on an index Z,

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

where Z is the average of d stocks,

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

and

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

with volatilities

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

The correlation between Wiener processes is given by

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

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

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

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

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

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

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

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

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

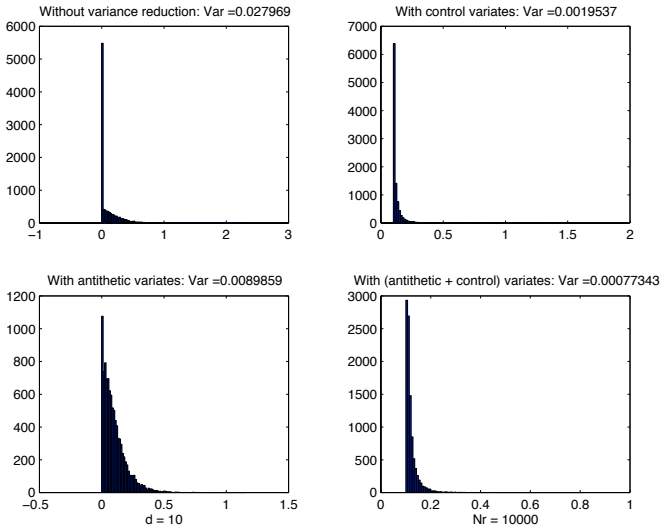
compute

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

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

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

(iv) Discuss the results from (i)-(iii). Does it pay to use variance reduction?


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

Chapter 6

Finite Difference Methods

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

6.1 American Options

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

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

An American put option is a contract that gives the possibility to sell a stock for a fixed price K up to time T. Therefore the derivation of option values in Chapter 4 shows that European and American options have the formulations:

(i) The price of an European put option is

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

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

$$f_A(t,s) \equiv \max_{t \le \tau \le T} E[\ e^{-r(\tau-t)} \max\big(K - S(\tau), 0\big)|\ S(t) = s\ ]. \tag{6.1}$$

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

$$f_t + rSf_S + \frac{1}{2}\sigma^2 S^2 f_{SS} = rf. \tag{6.2}$$

If, for a fixed stock price $s = S(T - \Delta t)$, there holds $f(T - \Delta t, s) < \max(K - s, 0)$ then keeping the option gives the expected value $f(T - \Delta t, s)$ which is clearly less than the value $\max(K - s, 0)$ obtained by selling at time $T - \Delta t$. Therefore it is optimal to sell if $f(T - \Delta t, s) < \max(K - s, 0) \equiv f_F$. Modify the initial data at $t = T - \Delta t$ to $\max(f(T - \Delta t, s), f_F)$ and repeat the step (6.2) for $(T - 2\Delta t, T - \Delta t)$ and so on. The price of the American option is obtained as the limit of this solution as $\Delta t \to 0$.

Example 6.1. A corresponding Monte Carlo method based on (6.1) requires simulation of expected values $E[e^{-r\tau} \max(K - S(\tau), 0)]$ for many different possible selling time strategies τ until an approximation of the maximum values is found. Since the τ need to depend on ω, with M time steps and N realizations there are M^N different strategies.

Note that the optimal selling strategy

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

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

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

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

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

(2) Use the Euler and central difference methods for the equation (6.2)

$$\begin{aligned} \partial_t f_A &\simeq \tfrac{\bar{f}_{n,i} - \hat{f}_{n-1,i}}{\Delta t} & \partial_S f_A &\simeq \tfrac{\bar{f}_{n,i+1} - \bar{f}_{n,i-1}}{2\Delta S} \\ \partial_{SS} f_A &\simeq \tfrac{\bar{f}_{n,i+1} - 2\bar{f}_{n,i} + \bar{f}_{n,i-1}}{(\Delta S)^2} & f_A &\simeq \bar{f}_{n,i}. \end{aligned}$$

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

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

(4) Compare the prediction with selling by letting

$$\bar{f}_{n-1,i} = \max\big(\hat{f}_{n-1,i}, \max(K - i\Delta S, 0)\big),$$

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

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

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

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

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

Let us now study general finite difference methods for partial differential equations. The motivation to introduce general finite difference methods in contrast to study only the binomial and trinomial tree methods is that higher order methods, such as the Crank-Nicolson method below, are more efficient to solve e.g. (6.2).

The error for the binomial and the trinomial tree method applied to the partial differential equation (6.2) for a European option is $\varepsilon = \mathcal{O}(\Delta t + (\Delta s)^2)$, which is clearly the same for the related forward and backward Euler methods. The work is then $\mathcal{A} = \mathcal{O}((\Delta t \Delta s)^{-1})$, so that $\mathcal{A} = \mathcal{O}(\varepsilon^{-3/2})$. For the Crank-Nicolsen method the accuracy is $\varepsilon = \mathcal{O}((\Delta t)^2 + (\Delta s)^2)$ and the work is still $\mathcal{A} = \mathcal{O}((\Delta t \Delta s)^{-1})$, which implies the improved bound $\mathcal{A} = \mathcal{O}(\varepsilon^{-1})$. For a general implicit method with a smooth exact solution in $[0, T] \times \mathbb{R}^d$ the accuracy is $\varepsilon = \mathcal{O}((\Delta t)^q + (\Delta s)^p)$ with the miminal work (using e.g. the multigrid method) $\mathcal{A} = \mathcal{O}(\frac{q^2}{\Delta t}(\frac{p^2}{\Delta s})^d)$, which gives $\mathcal{A} = \mathcal{O}(\frac{q^2}{\varepsilon^{1/q}}(\frac{p^2}{\varepsilon^{1/p}})^d)$. In the next section we derive these error estimates for some model problems.

6.2 Lax Equivalence Theorem

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

The ingredients of Lax Equivalence Theorem 6.4 are:

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

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

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

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

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

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

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

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

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

and consequently the convergence holds

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

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

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

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

$$\begin{aligned} &u'(t) = Au(t) \quad 0 < t < 1, \\ &u(0) = u_0. \end{aligned} \tag{6.3}$$

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

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

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

Then:

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

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

This stability and consistency imply the convergence

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

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

Example 6.6. Consider the heat equation

$$\begin{aligned} u_t &= u_{xx} \quad t > 0, \\ u(0) &= u_0. \end{aligned} \tag{6.4}$$

Verify the stability and consistency conditions in Lax Equivalence Theorem.

Solution. Apply the Fourier transform to equation (6.4),

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

so that

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

Therefore $\hat{H} = e^{-\Delta t \omega^2}$ is the exact solution operator for one time step, i.e. $\hat{u}(t + \Delta t) = \hat{H}\hat{u}(t)$. Consider the difference approximation of (6.4)

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

which shows

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

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

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

(i) We have

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

Therefore the condition

$$\|\hat{G}^n\|_{L^\infty} \leq e^{Kn\Delta t} \tag{6.5}$$

implies L^2-stability.

(ii) We have

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

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

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

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

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

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

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

Therefore the consistency condition reduces to

$$\begin{aligned} \| (\hat{G} - \hat{H})\hat{u}_0 \| &\le \|K\Delta t^2\omega^4\hat{u}_0\| \\ &\le K\Delta t^2\|\partial_{xxxx}u_0\|_{L^2}. \end{aligned} \tag{6.6}$$

(iii) The stability (6.5) holds if

$$\|\hat{G}\|_{L^\infty} \equiv \sup_\omega |\hat{G}(\omega)| = \max_\omega |1 - 2\lambda + 2\lambda\cos\Delta x\omega| \le 1, \tag{6.7}$$

which requires

$$\lambda = \frac{\Delta t}{\Delta x^2} \le \frac{1}{2}. \tag{6.8}$$

The L^2-stability condition (6.7) is called the von Neuman stability condition.

(iv) Convergence follows by the estimates (6.6), (6.7) and $\|\hat{H}\|_{L^\infty} \le 1$

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

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

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

$$u' + \lambda u = 0$$
$$u(0) = 1$$

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

(1) the exact solution satisfies

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

(2) the forward Euler method

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

(3) the backward Euler method

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

(4) the trapezoidal method

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

and

(5) the Lax-Wendroff method

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

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

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

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

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

Property (1) shows that for $\lambda > 0$ case 3 and 4 are unconditionally stable. However Property (2) and (3) refine this statement and imply that only case 3 has the same damping behavior for large λ as the exact solution. Although the damping Property (2) is not necessary to prove convergence it is advantegous to have for proplems with many time scales, e.g. for a system of equations (6.3) where A has eigenvalues $\lambda_i \leq 1,\ i = 1, \ldots, N$ and some $\lambda_j \ll -1$, (why?).

The unconditionally stable methods, e.g. case 3 and 4, are in general more efficient to solve parabolic problems, such as the Black-Scholes equation (6.2), since they require for the same accuracy fewer time steps than the explicit methods, e.g. case 2 and 5. Although the work in each time step for the unconditionally stable methods may be larger than for the explicit methods.

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

Chapter 7

The Finite Element Method and Lax-Milgram's Theorem

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

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

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

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

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

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

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

(1) an automatic mesh generator,

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

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

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

7.1 The Finite Element Method

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

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

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

(3) choice of a basis for V_h, and

(4) solution of the discrete system of equations.

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

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

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

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

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

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

where

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

Remark 7.1. The integration by parts in (7.2) shows that a smooth solution of equation (7.1) satisfies the variational formulation (7.3). For a solution of the variational formulation (7.3) to also be a solution of the equation (7.1), we need additional conditions

on the regularity of the functions a, r and f so that u'' is continuous. Then the following integration by parts yields, as in (7.2),

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

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

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

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

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

$$V_h = \{v \in V \quad : \quad v(x)\,|_{(x_i,x_{i+1})} = c_i x + d_i, \text{ i.e. } v \text{ is linear on } (x_i, x_{i+1}), i = 0, \cdots, N \\ \text{and } v \text{ is continuous on } (0,1)\}.$$

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

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

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

Step 3. *Choose a basis for V_h.*
Let us introduce the basis functions $\phi_i \in V_h$, for $i = 1, ..., N$, defined by

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

A function $v \in V_h$ has the representation

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

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

Step 4. *Solve the discrete problem (7.4).*
Using the basis functions ϕ_i, for $i = 1, ..., N$ from Step 3, we have

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

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

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

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

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

where

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

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

Example 7.2. Consider the following two dimensional problem,

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

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

(i) Variational formulation in the infinite dimensional space.

Let

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

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

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

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

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

where

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

(ii) Variational formulation in the finite dimensional space.

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

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

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

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

(iii) Choose a basis for V_0^h.

As in the one dimensional problem, choose the basis $\phi_j \in V_0^h$ such that

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

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

(iv) Solve the discrete system.

Let

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

Then (7.9) can be written in matrix form,

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

□

7.2 Error Estimates and Adaptivity

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

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

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

Theorem 7.3 (Lax-Milgram)**.** *Let V be a Hilbert space with norm $\|\cdot\|_V$ and scalar product $(\cdot,\cdot)_V$ and assume that A is a bilinear functional and L is a linear functional that satisfy:*

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

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

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

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

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

7.2.1 An A Priori Error Estimate

The approximation property of the space V_h can be characterized by

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

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

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

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

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

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

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

so that

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

which after summation of the intervals proves (7.11).

Next, we have

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

so by (7.12)

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

which after summation of the intervals proves the lemma. □

Our derivation of the a priori error estimate

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

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

(1) error representation based on the *ellipticity*

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

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

(2) the *orthogonality*

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

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

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

(3) the *continuity*

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

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

(4) the *interpolation estimates*

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

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

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

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

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

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

Therefore, by Step 1

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

which implies the a priori estimate

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

where $C = K(u)$. □

7.2.2 An A Posteriori Error Estimate

Example 7.5. Consider the model problem (7.1), namely,

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

Then

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

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

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

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

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

Lemma 7.6 implies

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

which also shows that

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

where $C = K'(u_h)$. □

Lemma 7.6. *There is a constant C, independent of u and u_h, such that,*

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

Exercise 7.7. Use the interpolation estimates in Lemma 7.4 to prove Lemma 7.6.

7.2.3 An Adaptive Algorithm

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

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

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

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

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

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

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

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

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

then go to Step 2.

7.3 Lax-Milgram's Theorem

Theorem 7.8. *Suppose A is symmetric, i.e. $A(u,v) = A(v,u) \ \ \forall u, v \in V$, then (Variational problem) $\Longleftrightarrow$ (Minimization problem) with*

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

where

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

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

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

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

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

Therefore

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

□

Theorem 7.9 (Lax-Milgram)**.** *Let V be a Hilbert space with norm $\|\cdot\|_V$ and scalar product $(\cdot,\cdot)_V$ and assume that A is a bilinear functional and L is a linear functional that satisfy:*

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

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

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

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

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

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

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

Let

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

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

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

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

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

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

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

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

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

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

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

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

so that

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

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

Example 7.10. Determine conditions for the functions k, r and $f : \Omega \to \mathbb{R}$ such that the assumptions in the Lax-Milgram theorem are satisfied for the following elliptic partial differential equation in $\Omega \subset \mathbf{R}^2$

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

Solution. This problem satisfies (Var) with

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

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

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

The ellipticity follows by

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

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

The continuity of A is a consequence of

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

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

Finally, the functional L is continuous, since

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

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

Example 7.11. Verify that the assumption of the Lax-Milgram theorem are satisfied for the following problem,

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

Solution. This problem satisfies (Var) with

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

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

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

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

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

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

provided $v(0) = 0$. Since

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

and by Cauchy's inequality

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

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

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

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

Chapter 8

Optimal Control and Inverse Problems

The purpose of Optimal Control is to influence the behavior of a dynamical system in order to achieve a desired goal. *Optimal control* has a large variety of applications where the dynamics can be controlled optimally, such as aerospace, aeronautics, chemical plants, mechanical systems, finance and economics, but also to solve *inverse problems* where the goal is to determine input data in an equation from its solution values. An important application we will study in several settings is to determine the "data" in differential equations models using optimally controlled reconstructions of measured "solution" values.

Inverse problems are typically harder to solve numerically than forward problems since they are often ill-posed (in contrast to forward problems), where ill-posed is the opposite of well-posed and a problem is defined to be well-posed if the following three properties holds

(1) there is a solution,

(2) the solution is unique, and

(3) the solution depends continuously on the data.

It is clear that a solution that does not depend continuously on its data is difficult to approximate accurately, since a tiny perturbation of the data (either as measurement error and/or as numerical approximation error) may give a large change in the solution. Therefore, the ill-posedness of inverse and optimal control problems means that they need to be somewhat modified to be solved: we call this to *regularize* the problem. Optimal control theory is suited to handle many inverse problems for differential equations, since we may formulate the objective – for instance to optimally reconstruct measured data or to find an optimal design – with the differential equation as a constraint. This chapter explains:

- the reason to regularize inverse problems in an optimal control setting,

- a method how to regularize the control problem, and
- in what sense the regularized problem approximates the original problem.

To give some intuition on optimal control and to introduce some basic concepts let us consider a hydro-power generator in a river. Suppose that we are the owners of such a generator, and that our goal is to maximise our profit by selling electricity in some local electricity market. This market will offer us buying prices at different hours, so one decision we have to make is when and how much electricity to generate. To make this decision may not be a trivial task, since besides economic considerations, we also have to meet technical constraints. For instance, the power generated is related to the amount of water in the reservoir, the turbined flow and other variables. Moreover, if we want a plan for a period longer than just a few days the water inflow to the lake may not be precisely known, making the problem stochastic.

We can state our problem in optimal control terms as the maximization of an *objective function*, the expected profit from selling electricity power during a given period, with respect to *control functions*, like the hourly turbined flow. Observe that the turbined flow is positive and smaller than a given maximum value, so it is natural to have a set of *feasible controls*, namely the set of those controls we can use in practice. In addition, our dynamical system evolves according to a given law, also called *the dynamics*, which here comes from a mass balance in the dam's lake. This law tells us how the *state variable*, the amount of water in the lake, evolves with time according to the control we give. Since the volume in the lake cannot be negative, there exist additional constraints, known as *state constraints*, that have to be fulfilled in the optimal control problem.

After introducing the formulation of an optimal control problem the next step is to find its solution. As we shall see, the optimal control is closely related with the solution of a nonlinear partial differential equation, known as the Hamilton-Jacobi-Bellman equation. To derive the Hamilton-Jacobi-Bellman equation we shall use the dynamic programming principle, which relates the solution of a given optimal control problem with solutions to simpler problems.

8.1 The Determinstic Optimal Control Setting

A mathematical setting for optimally controlling the solution to a deterministic ordinary differential equation

$$\begin{aligned} \dot{X}^s &= f(X^s, \alpha^s) \quad t < s < T \\ X^t &= x \end{aligned} \tag{8.1}$$

is to minimize

$$\inf_{\alpha \in \mathcal{A}} \Big(\int_t^T h(X^s, \alpha^s)\, \mathrm{d}s + g(X^T) \Big) \tag{8.2}$$

for given *cost functions* $h : \mathbb{R}^d \times [t, T] \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}$ and a given set of *control functions* $\mathcal{A} = \{\alpha : [t, T] \to A\}$ and *flux* $f : \mathbb{R}^d \times A \to \mathbb{R}^d$. Here A is a given compact subset of some $\mathbb{R}^m$.

8.1.1 Examples of Optimal Control

Example 8.1 (Optimal control of spacecraft)**.** To steer a spacecraft with minimal fuel consumption to an astronomical body may use the gravitational force from other bodies. The dynamics is determined by the classical Newton's laws with forces depending on the gravity on the spacecraft and its rocket forces, which is the control cf. [28].

Example 8.2 (Inverse problem: Parameter reconstruction)**.** The option values can be used to detemine the volatility function implicitly. The objective in the optimal control formulation is then to find a volatility function that yields option prices that deviate as little as possible from the measured option prices. The dynamics is the Black-Scholes equation with the volatility function to be determined, that is the dynamics is a determinstic partial differential equation and the volatility is the control function, see Section 8.2.1.1. This is a typical inverse problem: it is called inverse because in the standard view of the Black-Scholes equation relating the option values and the volaility, the option price is the unknown and the volatility is the data; while here the formulation is reversed with option prices as data and volatility as unknown in the same Black-Scholes equation.

Example 8.3 (Inverse problem: Weather prediction)**.** The incompressible Navier-Stokes equations are used to forecast weather. The standard mathematical setting of this equation is an initial value problem with unknown velocity and pressure to be determined from the initial data: in weather prediction one can use measured velocity and pressure not only at a single initial instance but data given over a whole time history. An optimal control formulation of the weather prediction is to find the first initial data (the control) matching the time history of measured velocity and pressure with the Navier-Stokes dynamics as constraint. Such an optimal control setting improves the accuracy and makes longer forecast possible as compared to the classical initial value problem, see [32], [**?**]. This is an inverse problem since the velocity and pressure are used to determine the "initial data".

Example 8.4 (Merton's stochastic portfolio problem)**.** A basic problem in finance is to choose how much to invest in stocks and in bonds to maximize a final utility function. The dynamics of the portfolio value is then stochastic and the objective is to maximize an expected value of a certain (utility) function of the portfolio value, see section 8.3.1.

Example 8.5 (Euler-Lagrange equation)**.** The shape of a soap bubble between a wire frame can be deterimined as the surface that minimizes the bubble area. For a surface in $\mathbb{R}^3$ described by $\{(x, u(x)) : x \in \Omega \subset \mathbb{R}^2\}$ the area is given by

$$\int_\Omega \sqrt{1 + |\nabla u|^2} dx.$$

Here the whole surface is the control function, and given a wire $\{(x, g(x)) : x \in \partial\Omega\}$, the minimal surface solves the Euler-Lagrange equation,

$$\begin{aligned} \operatorname{div}\left(\frac{\nabla u}{\sqrt{1 + |\nabla u|^2}}\right) &= 0, \quad \text{in } \Omega, \\ u &= g, \quad \text{on } \partial\Omega. \end{aligned}$$

Example 8.6 (Inverse problem: Optimal design)**.** An example of optimal design is to construct an electrical conductor to minimize the power loss by placing a given amount of conductor in a given domain, see Section 8.2.1.2. This is an inverse problem since the conductivity is determined from the electric potential in an equation where the standard setting is to determine the electric potential from the given conductivity.

8.1.2 Approximation of Optimal Control

Optimal control problems can be solved by the Lagrange principle or dynamic programming. The *dynamic programming* approach uses the value function, defined by

$$u(x,t) := \inf_{\alpha\in\mathcal{A}} \Big(\int_t^T h(X^s,\alpha^s)\,\mathrm{d}s + g(X^T) \Big), \tag{8.3}$$

for the ordinary differential equation (8.1) with $X_t \in \mathbb{R}^d$, and leads to solution of a non linear *Hamilton-Jacobi-Bellman* partial differential equation

$$\begin{aligned} &\partial_t u(x,t) + \underbrace{\min_{\alpha\in A} \big(f(x,\alpha)\cdot\partial_x u(x,t) + h(x,\alpha) \big)}_{H(\partial_x u(x,t),x)} = 0, \quad t<T, \\ &u(\cdot,T) = g, \end{aligned} \tag{8.4}$$

in $(x,t) \in \mathbb{R}^d \times \mathbb{R}_+$. The *Lagrange principle* (which seeks a minimum of the cost with the dynamics as a constraint) leads to the solution of a Hamiltonian system of ordinary differential equations, which are the characteristics of the Hamilton-Jacobi-Bellman equation

$$\begin{aligned} X'^t &= f(X^t,\alpha^t), \quad X_0 \text{ given}, \\ -\lambda_i'^t &= \partial_{x_i} f(X^t,\alpha^t)\cdot\lambda^t + \partial_{x_i} h(X^t,\alpha^t), \quad \lambda^T = g'(X^T), \\ \alpha^t &\in \operatorname{argmin}_{a\in A} \Big(\lambda^t\cdot f(X^t,a) + h(X^t,a) \Big), \end{aligned} \tag{8.5}$$

based on the Pontryagin Principle. The next sections explain these two methods.

The non linear Hamilton-Jacobi partial differential approach has the theoretical advantage of well established theory and that a global minimum is found; its fundamental drawback is that it cannot be used computationally in high dimension $d \gg 1$, since the computational work increases exponentially with the dimension d. The Lagrange principle has the computational advantage that high dimensional problems, $d \gg 1$, can often be solved and its drawback is that in practice only local minima can be found computationally, often with some additional error introduced by a regularization method. Another drawback with the Lagrange principle is that it (so far) has no efficient implementation in the natural stochastic setting with adapted Markov controls, while the Hamilton-Jacobi PDE approach directly extends to such stochastic controls, see Section 8.3; as a consequence computations of stochastic controls is basically limited to low dimensional problems.

8.1.3 Motivation of the Lagrange formulation

Let us first review the Lagrange multiplier method to minimize a function subject to a constraint $\min_{x\in A,\, y=g(x)} F(x,y)$. Assume $F : \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}$ is a differentiable function. The goal is to find the minimum $\min_{x\in A} F(x, g(x))$ for a given differentiable function $g : \mathbb{R}^d \to \mathbb{R}^n$ and a compact set $A \subset \mathbb{R}^d$. This problem leads to the usual necessary condition for an interior minimum

$$\frac{\mathrm{d}}{\mathrm{d}x} F\big(x, g(x)\big) = \partial_x F\big(x, g(x)\big) + \partial_y F\big(x, g(x)\big)\partial_x g(x) = 0. \tag{8.6}$$

An alternative method to find the solution is to introduce the Lagrangian function $\mathcal{L}(\lambda, y, x) := F(x, y) + \lambda \cdot \big(y - g(x)\big)$ with the Lagrange multiplier $\lambda \in \mathbb{R}^n$ and choose λ appropriately to write the necessary condition for an interior minimum

$$\begin{aligned} 0 &= \partial_\lambda \mathcal{L}(\lambda, y, x) = y - g(x), \\ 0 &= \partial_y \mathcal{L}(\lambda, y, x) = \partial_y F(x, y) + \lambda, \\ 0 &= \partial_x \mathcal{L}(\lambda, y, x) = \partial_x F(x, y) - \lambda \cdot \partial_x g(x). \end{aligned}$$

Note that the first equation is precisely the constraint. The second equation determines the multiplier to be $\lambda = -\partial_y F(x, y)$. The third equation yields for this multiplier $\partial_x \mathcal{L}(-\partial_y F(x, y), y, x) = \frac{d}{dx} F\big(x, g(x)\big)$, that is the multiplier is chosen precisely so that the partial derivative with respect to x of the Lagrangian is the total derivative of the objective function $F\big(x, g(x)\big)$ to be minimized. This Lagrange principle is often practical to use when the constraint is given implicitly, e.g. as $g(x, y) = 0$ with a differentiable $g : \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}^n$; then the condition $\det \partial_y g(x, y) \neq 0$ in the implicit function theorem implies that the function $y(x)$ is well defined and satisfies $g\big(x, y(x)\big) = 0$ and $\partial_x y = -\partial_y g(x, y)^{-1} \partial_x g(x, y)$, so that the Lagrange multiplier method works.

The Lagrange principle for the optimal control problem (8.1) -(8.2), to minimize the cost with the dynamics as a constraint, leads to the Lagrangian

$$\mathcal{L}(\lambda, X, \alpha) := g(X^T) + \int_0^T h(X^s, \alpha^s)\,\mathrm{d}s + \int_0^T \lambda^s \cdot \big(f(X^s, \alpha^s) - \dot{X}\big)\,\mathrm{d}s \tag{8.7}$$

with a Lagrange multiplier function $\lambda : [0, T] \to \mathbb{R}^d$. Differentiability of the Lagrangian leads to the necessary conditions for a constrained interior minimum

$$\begin{aligned} \partial_\lambda \mathcal{L}(X, \lambda, \alpha) &= 0, \\ \partial_X \mathcal{L}(X, \lambda, \alpha) &= 0, \\ \partial_\alpha \mathcal{L}(X, \lambda, \alpha) &= 0. \end{aligned} \tag{8.8}$$

Our next step is to verify that the two first equations above are the same as the two first in (8.5) and that the last equation is implied by the stronger Pontryagin principle in the last equation in (8.5). We will later use the Hamilton-Jacobi equation in the dynamic programming approach to verify the Pontryagin principle.

The first equation. Choose a real valued continuous function $v : [0,T] \to \mathbb{R}^d$ and define the function $L : \mathbb{R} \to \mathbb{R}$ by $L(\epsilon) := \mathcal{L}(X, \lambda + \epsilon v, \alpha)$. Then the first of the three equations means precisely that $L'(0) = \frac{d}{d\epsilon}\mathcal{L}(X, \lambda + \epsilon v, \alpha)|_{\epsilon=0} = 0$, which implies that

$$0 = \int_0^T v^s \cdot \left(f(X^s, \alpha^s) - \dot{X}^s\right) \mathrm{d}s$$

for any continuous function v. If we assume that $f(X^s, \alpha^s) - \dot{X}^s$ is continuous we obtain $f(X^s, \alpha^s) - \dot{X}^s = 0$: since if $\beta(s) := f(X^s, \alpha^s) - \dot{X}^s \neq 0$ for some s there is an interval where β is either positive or negative; by choosing v to be zero outside this interval we conclude that β is zero everywhere and we have derived the first equation in (8.5).

The second equation. The next equation $\frac{d}{d\epsilon}\mathcal{L}(X + \epsilon v, \lambda, \alpha)|_{\epsilon=0} = 0$ needs $v^0 = 0$ by the initial condition on X^0 and leads by integration by parts to

$$\begin{aligned}
0 &= \int_0^T \lambda^s \cdot \left(\partial_{X_i} f(X^s, \alpha^s) v_i^s - \dot{v}^s\right) + \partial_{X_i} h(X^s, \alpha^s) v_i^s \,\mathrm{d}s + \partial_{X_i} g(X^T) v_i^T \\
&= \int_0^T \lambda^s \cdot \partial_{X_i} f(X^s, \alpha^s) v_i^s + \dot{\lambda} \cdot v^s + \partial_{X_i} h(X^s, \alpha^s) v_i^s \,\mathrm{d}s \\
&\quad + \lambda^0 \cdot \underbrace{v^0}_{=0} - \left(\lambda^T - \partial_X g(X^T)\right) \cdot v^T \\
&= \int_0^T \left(\partial_X f^*(X^s, \alpha^s)\lambda^s + \dot{\lambda}^s + \partial_X h(X^s, \alpha^s)\right) \cdot v^s \,\mathrm{d}s \\
&\quad - \left(\lambda^T - \partial_X g(X^T)\right) \cdot v^T,
\end{aligned}$$

using the summation convention $a_i b_i := \sum_i a_i b_i$. Choose now the function v to be zero outside an interior interval where possibly $\partial_X f^*(X^s, \alpha^s)\lambda^s + \dot{\lambda}^s + \partial_X h(X^s, \alpha^s)$ is non zero, so that in particular $v^T = 0$. We see then that in fact $\partial_X f^*(X^s, \alpha^s)\lambda^s + \dot{\lambda}^s + \partial_X h(X^s, \alpha^s)$ must be zero (as for the first equation) and we obtain the second equation in (8.5). Since the integral in the right hand side vanishes, varying v^T shows that the final condition for the Lagrange multiplier $\lambda^T - \partial_X g(X^T) = 0$ also holds.

The third equation. The third equation in (8.8) implies as above that for any function $v(t)$ compactly supported in A

$$0 = \int_0^T \lambda^s \cdot \partial_\alpha f(X^s, \alpha^s) v + \partial_\alpha h(X^s, \alpha^s) v \,\mathrm{d}s$$

which yields

$$\lambda^s \cdot \partial_\alpha f(X^s, \alpha^s) + \partial_\alpha h(X^s, \alpha^s) = 0 \tag{8.9}$$

in the interior $\alpha \in A - \partial A$ minimum point (X, λ, α). The last equation in (8.5) is a stronger condition: it says that α is a minimizer of $\lambda^s \cdot f(X^s, a) + h(X^s, a) = 0$ with respect to $a \in A$, which clearly implies (8.9) for interior points $\alpha \in A - \partial A$. To derive the Pontryagin principle we will use dynamic programming and the Hamilton-Jacobi-Bellman equation which is the subject of the next section.

8.1.4 Dynamic Programming and the Hamilton-Jacobi-Bellman Equation

The dynamic programming view to solve optimal control problems is based on the idea to track the optimal solution backwards: at the final time the value function is given $u(x,T) = g(x)$ and then, recursively for small time step backwards, find the optimal control to go from each point (x,t) on the time level t to the time level $t+\Delta t$ with the value function $u(\cdot, t+\Delta t)$, see Figure 8.1. Assume for simplicity first that $h \equiv 0$ then any path $X : [t, t+\Delta t] \to \mathbb{R}^d$ starting in $X^t = x$ will satisfy

$$u(x,t) = \inf_{\alpha:[t,t+\Delta t]\to A} u(X^{t+\Delta t}, t+\Delta t),$$

so that if u is differentiable

$$\mathrm{d}u(X^t,t) = \big(\partial_t u(X^t,t) + \partial_x u(X^t,t)\cdot f(X^t,\alpha_t)\big)\,\mathrm{d}t \geq 0, \tag{8.10}$$

since a path from (x,t) with value $u(x,t)$ can lead only to values $u(X^{t+\Delta t}, t+\Delta t)$ which are not smaller than $u(x,t)$. If also the infimum is attained, then an optimal path X^t_* exists, with control α^t_*, and satisfies

$$\mathrm{d}u(X^t_*,t) = \big(\partial_t u(X^t_*,t) + \partial_x u(X^t_*,t)\cdot f(X^t_*,\alpha^t_*)\big)\,\mathrm{d}t = 0. \tag{8.11}$$

The combination of (8.10) and (8.11) implies that

$$\partial_t u(x,t) + \min_{\alpha\in A}\big(\partial_x u(x,t)\cdot f(x,\alpha)\big) = 0 \quad t < T$$
$$u(\cdot,T) = g,$$

which is the Hamilton-Jacobi-Bellman equation in the special case $h \equiv 0$.

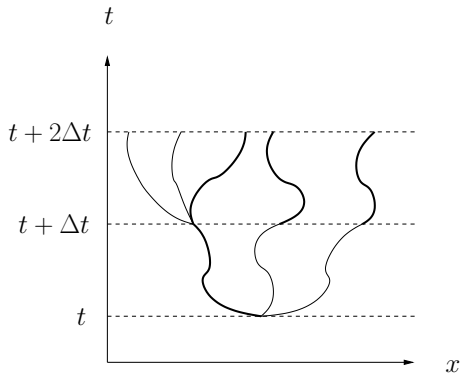


Figure 8.1: Illustration of dynamics programming.

The case with h non zero follows similarly by noting that now

$$0 = \inf_{\alpha:[t,t+\Delta t]\to A}\Big(\int_t^{t+\Delta t} h(X^s,\alpha^s)\,\mathrm{d}s + u(X^{t+\Delta t}, t+\Delta t) - u(x,t)\Big), \tag{8.12}$$

which for differentiable u implies the Hamilton-Jacobi-Bellman equation (8.4)

$$\begin{aligned} 0 &= \inf_{\alpha\in A}\big(h(x,\alpha)+\partial_t u(x,t)+\partial_x u(x,t)\cdot f(x,\alpha)\big)\\ &= \partial_t u(x,t)+\underbrace{\min_{\alpha\in A}\big(\partial_x u(x,t)\cdot f(x,\alpha)+h(x,\alpha)\big)}_{=:H\left(\partial_x u(x,t),x\right)} \quad t<T,\\ g &= u(\cdot,T). \end{aligned}$$

Note that this derivation did not assume that an optimal path is attained, but that u is differentiable which in general is not true. There is fortunately a complete theory for non differentiable solutions to Hamilton-Jacobi equations, with its basics presented in Section 8.1.6. First we shall relate the Lagrange multiplier method with the Pontryagin principle to the Hamilton-Jacobi-Bellman equation using charateristics.

8.1.5 Characteristics and the Pontryagin Principle

The following theorem shows that the characteristics of the Hamilton-Jacobi equation is a Hamiltonian system.

Theorem 8.7. *Assume $u\in\mathcal{C}^2$, $H\in\mathcal{C}^1$ and*

$$\dot{X}^t=\partial_\lambda H\big(\lambda^t,X^t\big)$$

with $\lambda^t:=\partial_x u(X^t,t)$. Then the characteristics (X^t,λ^t) satisfy the Hamiltonian system

$$\begin{aligned} \dot{X}^t &= \partial_\lambda H(\lambda^t,X^t)\\ \dot{\lambda}^t &= -\partial_X H(\lambda^t,X^t). \end{aligned} \tag{8.13}$$

Proof. The goal is to verify that the construction of X^t implies that λ has the dynamics (8.13). The definition $\dot{X}^t=\partial_\lambda H(\lambda^t,X^t)$ implies by x-differentiation of the Hamilton-Jacobi equation along the path (X^t,t)

$$\begin{aligned} 0 &= \partial_{x_k}\partial_t u(X^t,t)+\sum_j \partial_{\lambda_j}H\big(\partial_x u(X^t,t),X^t\big)\underbrace{\partial_{x_k}\partial_{x_j}u(X^t,t)}_{=\partial_{x_j}\partial_{x_k}u}+\partial_{x_k}H\big(\partial_x u(X^t,t),X^t\big)\\ &= \frac{\mathrm{d}}{\mathrm{d}t}\partial_{x_k}u(X^t,t)+\partial_{x_k}H\big(\partial_x u(X^t,t),X^t\big) \end{aligned}$$

which by the definition $\lambda^t:=\partial_x u(X^t,t)$ is precisely $\dot{\lambda}^t+\partial_x H(\lambda^t,X^t)=0$. □

The next step is to relate the characteristics X^t,λ^t to the solution of the Lagrange principle (8.5). But note first that the Hamiltonian H in general is not differentiable, even if f and h are very regular: for instance $\dot{X}=f(X^t)$ and $h(x,\alpha)=x\alpha$ implies for $A=[-1,1]$ that the Hamiltonian becomes $H(\lambda,x)=\lambda f(x)-|x|$ which is only Lipschitz continuous, that is $|H(\lambda,x)-H(\lambda,y)|\le K|x-y|$ with the Lipschitz constant $K=1+\|\lambda\cdot\partial_x f(\cdot)\|_\infty$ in this case. In fact if f and h are bounded differentiable functions the Hamiltonian will always be Lipschitz continuous satisfying $|H(\lambda,x)-H(\nu,y)|\le K(|\lambda-\nu|+|x-y|)$ for some constant K, see Exercise **??**.

Theorem 8.8. *Assume that f, h are x-differentiable in (x, α_*) and a control α_* is optimal for a point (x, λ), i.e.*

$$\lambda \cdot f(x, \alpha_*) + h(x, \alpha_*) = H(\lambda, x),$$

and suppose also that H is differentiable in the point or that α_* is unique. Then

$$\begin{aligned} f(x, \alpha_*) &= \partial_\lambda H(\lambda, x), \\ \lambda \cdot \partial_{x_i} f(x, \alpha_*) + \partial_{x_i} h(x, \alpha_*) &= \partial_{x_i} H(\lambda, x). \end{aligned} \tag{8.14}$$

Proof. We have for any $w, v \in \mathbb{R}^d$

$$\begin{aligned} H(\lambda + w, x + v) - H(\lambda, x) &\leq (\lambda + w) \cdot f(x + v, \alpha_*) + h(x + v, \alpha_*) \\ &\quad - \lambda \cdot f(x, \alpha_*) - h(x, \alpha_*) \\ &= w \cdot f(x, \alpha_*) + \sum_{i=1}^{d} (\lambda \cdot \partial_{x_i} f + \partial_{x_i} h) v_i + o(|v| + |w|) \end{aligned}$$

which implies (8.14) by choosing w and v in all directions. □

This Theorem shows that the Hamiltonian system (8.13) is the same as the system (8.5), given by the Lagrange principle using the optimal control α_* with the Pontryagin principle

$$\lambda \cdot f(x, \alpha_*) + h(x, \alpha_*) = \inf_{\alpha \in A} \big(\lambda \cdot f(x, \alpha) + h(x, \alpha)\big) =: H(\lambda, x).$$

If α_* is not unique (i.e not a single point) the proof shows that (8.14) still holds for the optimal controls, so that $\partial_\lambda H$ and $\partial_x H$ become set valued. We conclude that non unique local controls α_* is the phenomenon that makes the Hamiltonian non differentiable in certain points. In particular a differentiable Hamiltonian gives unique optimal control fluxes $\partial_\lambda H$ and $\partial_x H$, even if α_* is not a single point. If the Hamiltonian can be explicitly formulated, it is therefore often practical to use the Hamiltonain system formulation with the variables X and λ, avoiding the control variable.

Clearly, the Hamiltonian needs to be differentiable for the Hamiltonian system to make sense; in fact its flux $(\partial_\lambda H, -\partial_x H)$ must be Lipschitz continuous to give well posedness. On the other hand we shall see that the Hamilton-Jacobi-Bellman formulation, based on dynamic programming, leads to non differentiable value functions u, so that classical solutions lack well posedness. The mathematical setting for optimal control therefore seemed somewhat troublesome both on the Hamilton-Jacobi PDE level and on the Hamilton ODE level. In the 1980's the situation changed: Crandall-Lions-Evans [8] formulated a complete well posedness theory for generalized so called viscosity solutions to Hamilton-Jacobi partial differential equations, allowing Lipschitz continuous Hamiltonians. The theory of viscosity solutions for Hamilton-Jacobi-Bellman partial differential equations provides good theoretical foundation also for non smooth controls.

In particular this mathematical theory removes one of Pontryagin's two reasons[1], but not the other, to favor the ODE approach (8.5) and (8.13): the mathematical theory of viscosity solutions handles elegantly the inherent non smoothness in control problems; analogous theoretical convergence results for an ODE approach was developed later based on the so called minmax solutions, see [38]; we will use an alternative ODE method to solve optimal control problems numerically based on regularized Hamiltonians, where we approximate the Hamiltonian with a two times differentiable Hamiltonian, see Section 8.2.

Before we formulate the generalized solutions, we show that classical solutions only exist for short time in general.

Example 8.9. The Hamilton-Jacobi equation

$$\partial_t u - \frac{1}{2}(\partial_x u)^2 = 0$$

has the characteristics

$$\dot{X}^t = -\lambda^t$$
$$\dot{\lambda}^t = 0,$$

which implies $\dot{X}^t =$ constant. If the initial data $u(\cdot, T)$ is a concave function (e.g. a smooth version of $-|x|$) characteristics X will collide, see Figure 8.2. We can understand this precisely by studying blow-up of the derivative w of $\partial_x u =: v$; since v satisfies

$$\partial_t v - \underbrace{\frac{1}{2}\partial_x(v^2)}_{v\partial_x v} = 0$$

we have by $x-$differentiation

$$\underbrace{\partial_t w - v\partial_x w}_{\frac{d}{dt}w(X^t,t)} - w^2 = 0,$$

which reduces to the ordinary differential equation for $z^t := w(X^t, t)$

$$\frac{\mathrm{d}}{\mathrm{d}t}z(t) = z^2(t).$$

Its separation of variables solution $\mathrm{d}z/z^2 = \mathrm{d}t$ yields $-1/z^t = t + C$. The constant becomes $C = -T - 1/z^T$, so that $z^t = 1/(t - T - 1/z^T)$ blows up to infinity at time $T - t = 1/z^T$. For instance if $z^T = -10$, the time to blow-up time is $1/10$.

[1] citation from chapter one in [33] "This equation of Bellman's yields an approach to the solution of the optimal control problem which is closely connected with, but different from, the approach described in this book (see Chapter 9). It is worth mentioning that the assumption regarding the continuous differentiability of the functional (9.8) [(8.3) here] is not fulfilled in even the simplest cases, so that Bellman's arguments yield a good heuristic method rather than a mathematical solution of the problem. The maximum principle, apart from its sound mathematical basis, also has the advantage that it leads to a system of ordinary differential equations, whereas Bellman's approach requires the solution of a partial differential equation."

8.1.6 Generalized Viscosity Solutions of Hamilton-Jacobi-Bellman Equations

Example 8.9 shows that Hamilton-Jacobi equations do in general not have global classical solutions – after finite time the derivative can become infinitely large even with smooth initial data and a smooth Hamiltonian. Therefore a more general solution concept is needed. We shall describe the so called viscosity solutions introduced by Crandall and Lions in [9], which can be characterised by the limit of viscous approximations u^ϵ satisfying for $\epsilon > 0$

$$\begin{aligned} \partial_t u^\epsilon(x,t) + H\big(\partial_x u^\epsilon(x,t), x\big) + \epsilon \partial_{xx} u^\epsilon(x,t) = 0 \quad t < T \\ u^\epsilon(\cdot, T) = g. \end{aligned}$$

The function u^ϵ is also a value function, now for the stochastic optimal control problem

$$dX^t = f(X^t, \alpha^t)dt + \sqrt{2\epsilon}\, dW^t \quad t > 0$$

with the objective to minimize

$$\min_\alpha \mathbb{E}\Big[g(X^T) + \int_0^T h(X^t, \alpha^t)dt \;\Big|\; X^0 \text{given}\Big],$$

over adapted controls $\alpha : [0,T] \to A$, where $W : [0,\infty) \to \mathbb{R}^d$ is the d-dimensional Wiener process with independent components. Here adapted controls means that α_t does not use values of W^s for $s > t$. Section 8.3 shows that the value function for this optimal control problem solves the second order Hamilton-Jacobi equation, that is

$$u^\epsilon(x,t) = \min_\alpha \mathbb{E}\Big[g(X^T) + \int_0^T h(X^t, \alpha^t)\, \mathrm{d}t \;\Big|\; X^t = x\Big].$$

Theorem 8.10 (Crandall-Lions)**.** *Assume f, h and g are Lipschitz continuous and bounded, then the limit $\lim_{\epsilon \to 0+} u^\epsilon$ exists. This limit is called the viscosity solution of the Hamilton-Jacobi equation*

$$\begin{aligned} \partial_t u(x,t) + H\big(\partial_x u(x,t), x\big) = 0 \quad t < T \\ u(\cdot, T) = g. \end{aligned} \tag{8.15}$$

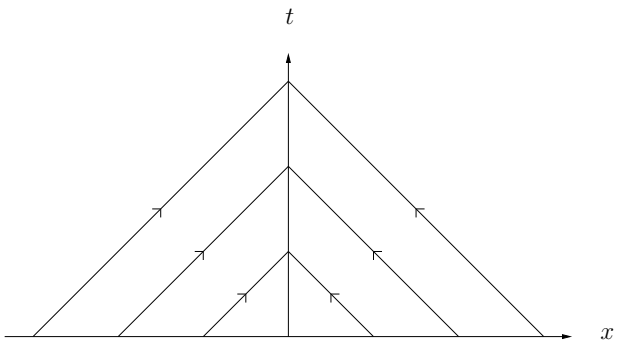


Figure 8.2: Characteristic curves colliding.

There are several equivalent ways to describe the viscosity solution directly without using viscous or stochastic approximations. We shall use the one based on sub and super differentials presented first in [8]. To simplify the notation introduce first the space-time coordinate $y = (x, t)$, the space-time gradient $p = (p_x, p_t) \in \mathbb{R}^{d+1}$ (related to $(\partial_x u(y), \partial_t u(y))$) and write the Hamilton-Jacobi operator $F(p, y) := p_t + H(p_x, x)$. For a bounded uniformly continuous function $v : \mathbb{R}^d \times [0, T] \to \mathbb{R}$ define for each space-time point y its sub differential set

$$\mathrm{D}^- v(y) = \{p \in \mathbb{R}^{d+1} : \liminf_{z \to 0} |z|^{-1} \big(v(y+z) - v(y) - p \cdot z\big) \geq 0\}$$

and its super differential set

$$\mathrm{D}^+ v(y) = \{p \in \mathbb{R}^{d+1} : \limsup_{z \to 0} |z|^{-1} \big(v(y+z) - v(y) - p \cdot z\big) \leq 0\}.$$

These two sets always exist (one may be empty), see Example 8.11; they degenerate to a single point, the space-time gradient of v, precisely if v is differentiable, that is when

$$\mathrm{D}^- v(y) = \mathrm{D}^+ v(y) = \{p\} \iff v(y+z) - v(y) - p \cdot z = o(z).$$

Example 8.11. Let $u(x) = -|x|$, then

$$\begin{aligned} &\mathrm{D}^+ u(x) = \mathrm{D}^- u(x) = \{-\mathrm{sgn(x)}\} \quad \mathrm{x} \neq 0 \\ &\mathrm{D}^- u(0) = \emptyset \quad x = 0 \\ &\mathrm{D}^+ u(0) = [-1, 1] \quad x = 0 \end{aligned}$$

see Figure 8.3.

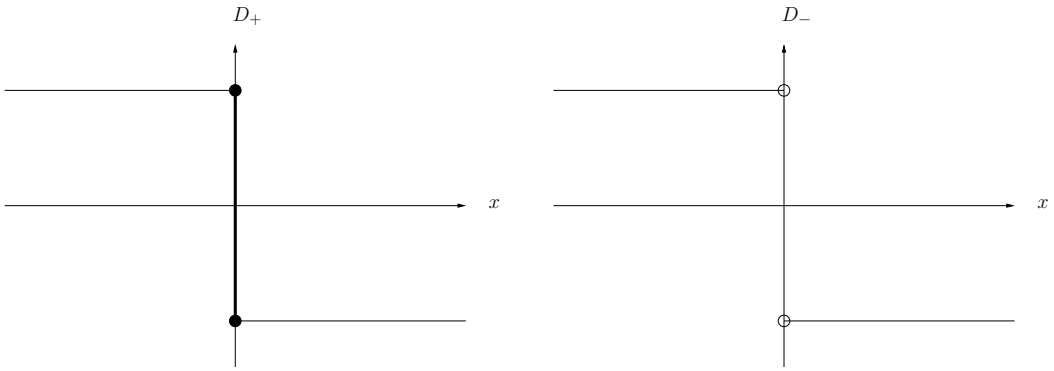


Figure 8.3: Illustration of the super and subdifferential sets for $-|x|$.

Definition 8.12 (Viscosity solution)**.** A bounded uniformly continuous function u is a viscosity solution to (8.15) if $u(\cdot, T) = g$ and for each point $y = (x, t)$

$$F(p, y) \geq 0 \quad \text{for all } p \in \mathrm{D}^+ u(y)$$

and

$$F(p, y) \leq 0 \quad \text{for all } p \in \mathrm{D}^- u(y).$$

Theorem 8.13. *The first variation of the value function is in the superdifferential.*

Proof. Consider an optimal path X_*, starting in $\bar{y} = (\bar{x}, \bar{t})$, with control α_*. We define the first variation, $(\lambda^{\bar{t}}, \nu^{\bar{t}}) \in \mathbb{R}^d \times \mathbb{R}$, of the value function along *this path*, with respect to perturbations in the initial point $\bar{y}$: let X_y be a path starting from a point $y = (x, t)$, close to $\bar{y}$, using the control α_*, the differentiability of the flux f and the cost h implies that the first variation satisfies

$$\lambda_i^{\bar{t}} = \lim_{z \to 0} z^{-1} \Big(\int_{\bar{t}}^{T} h(X^t_{\bar{x}+ze_i}, \alpha_*^t) - h(X^t_{\bar{x}}, \alpha_*^t) \, \mathrm{d}t + g(X^T_{\bar{x}+ze_i}) - g(X^T_{\bar{x}}) \Big) \tag{8.16}$$

and

$$\begin{aligned} -\dot{\lambda}^t &= \partial_X f(X_*^t, \alpha_*^t)\lambda^t + \partial_X h(X_*^t, \alpha_*^t) \quad \bar{t} < t < T, \\ \lambda^T &= g'(X_*^T), \end{aligned}$$

where e_i is the ith unit basis vector in $\mathbb{R}^d$. The definition of the value function shows that

$$-h(X_*^t, \alpha_*^t) = \frac{\mathrm{d}u}{\mathrm{d}t}(X_*^t, t) = \lambda^t \cdot f(X_*^t, \alpha_*^t) + \nu^t$$

so that

$$\nu^t = -\lambda^t \cdot f(X_*^t, \alpha_*^t) - h(X_*^t, \alpha_*^t).$$

Since the value function is the minimum possible cost, we have by (8.16)

$$\begin{aligned} &\limsup_{s \to 0+} s^{-1} \Big(u\big(\bar{y} + s(y - \bar{y})\big) - u(\bar{y}) \Big) \\ &\leq \limsup_{s \to 0+} s^{-1} \Big(\int_{\bar{t}}^{T} h(X^t_{\bar{y}+s(y-\bar{y})}, \alpha_*^t) \, \mathrm{d}t + g(X^T_{\bar{y}+s(y-\bar{y})}) \\ &\quad - \int_{\bar{t}}^{T} h(X^t_{\bar{y}}, \alpha_*^t) \, \mathrm{d}t + g(X^T_{\bar{y}}) \Big) \\ &= \Big(\lambda^t, -\big(\lambda^t \cdot f(X_*^t, \alpha_*^t) + h(X_*^t, \alpha_*^t)\big) \Big) \cdot (y - \bar{y}), \end{aligned}$$

which means precisely that the first variation is in the superdifferential. □

Theorem 8.14. *The value function is semi-concave, that is for any point (x, t) either the value function is differentiable or the sub differential is empty (i.e. $\mathrm{D}^- u(x, t) = \emptyset$ and $\mathrm{D}^+ u(x, t)$ is non empty).*

Proof. Assume that the subdifferential $\mathrm{D}^- u(y)$ has at least two elements p_- and p_+ (we will show that this leads to a contradiction). Then u is larger or equal to the wedge like function

$$u(y) \geq u(\bar{y}) + \max\big(p_- \cdot (y - \bar{y}), p_+ \cdot (y - \bar{y})\big), \tag{8.17}$$

see Figure 8.4. The definition of the value function shows that the right derivative satisfies

$$\limsup_{s \to 0+} s^{-1} \Big(u\big(\bar{y} + s(y - \bar{y})\big) - u(\bar{y}) \Big) \leq (\lambda, \nu) \cdot (y - \bar{y}) \tag{8.18}$$

where (λ, ν) is the first variation (in x and t) of u around the optimal path starting in $\bar{y}$. The wedge bound (8.17) implies

$$\limsup_{s \to 0+} s^{-1} \Big(u\big(\bar{y} + s(y - \bar{y})\big) - u(\bar{y}) \Big) \geq \max\big(p_- \cdot (y - \bar{y}), p_+ \cdot (y - \bar{y})\big),$$

but the value function cannot be both below a (λ, ν)-half plane (8.18) and above such wedge function, see Figure 8.5. Therefore the subdifferential can contain at most one point: either the subdifferential is empty or there is precisely one point p in the subdifferential and in this case we see that the the first variation coincides with this point $(\lambda, \nu) = p$, that is the value function is differentiable □

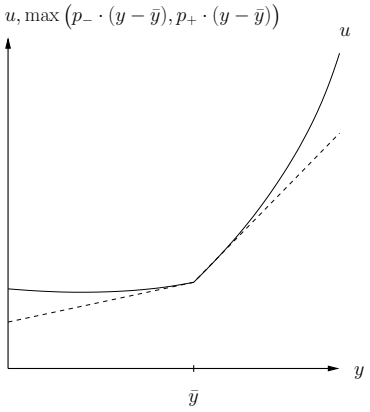


Figure 8.4: Characteristic curves colliding.

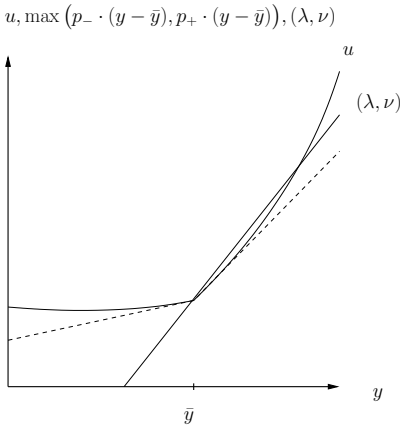


Figure 8.5: Characteristic curves colliding.

Theorem 8.15. *The value function is a viscosity solution.*

Proof. We have seen in Section 8.1.4 that for the points where the value function is differentiable it satisfies the Hamilton-Jacobi-Bellman equation. Theorem 8.14 shows that the value function u is semi-concave. Therefore, by Definition 8.12, it is enough to verify that $p \in \mathrm{D}^+u(x,t)$ implies $p_t + H(p_x, x) \geq 0$. Assume for simplicity that $h \equiv 0$.

There is a $p \in D^+u(x,t)$, which is the first variation of u along an optimal path (X^*, α^*), such that

$$\begin{aligned} p_t + H(p_x, x) &= p \cdot \big(f(x,\alpha), 1\big) \\ &\geq \limsup_{\Delta t \to 0+} \frac{u(X^{t+\Delta t}, t+\Delta t) - u(X^t, t)}{\Delta t} = 0, \end{aligned}$$

using the definition of the superdifferential and dynamic programming. This means that any optimal control yields a super differential point p satisfying $p_t + H(p_x, x) \geq 0$. To finish the proof we note that any point in the super differential set can for some $s \in [0,1]$ be written as a convex combination $sp^1 + (1-s)p^2$ of two points p^1 and p^2 in the super differential that correspond to (different) optimal controls. Since H is concave in p (see Exercise 8.19) there holds

$$\begin{aligned} & sp_t^1 + (1-s)p_t^2 + H\big(sp_x^1 + (1-s)p_x^2, x\big) \\ & \geq s\big(p_t^1 + H(p_x^1, x)\big) + (1-s)\big(p_t^2 + H(p_x^2, x)\big) \\ & \geq 0 \end{aligned}$$

which shows that u is a viscosity solution. The general case with non zero h is similar as in (8.12). □

Theorem 8.16. *Bounded uniformly continuous viscosity solutions are unique.*

The standard uniqueness proof uses a special somewhat complex doubling of variables technique, see [16] inspired by Kruzkov. The maximum norm stability of semi-concave viscosity solutions in Section 8.1.7 also implies uniqueness.

Example 8.17. Consider the function $u(x,t) = -|x|$. We have from Example 8.11

$$D^+u(x,t) = \begin{cases} (-\mathrm{sgn}(x), 0) & x \neq 0 \\ ([-1,1], 0) & x = 0 \end{cases}$$

and

$$D^-u(x,t) = \begin{cases} (-\mathrm{sgn}(x), 0) & x \neq 0 \\ \emptyset & x = 0. \end{cases}$$

Consequently for $H(\lambda, x) := (1 - |\lambda|^2)/2$ we obtain

$$\begin{aligned} p_t + H(p_x, x) \geq 0 \quad q \in D^+u(x,t) \\ p_t + H(p_x, x) = 0 \quad q \in D^-u(x,t) \end{aligned}$$

so that $-|x|$ is a viscosity solution to $\partial_t u + H(\partial_x u, x) = 0$. Similarly the function $u(x,t) = |x|$ satisfies

$$D^-u(x,t) = \begin{cases} (\mathrm{sgn}(x), 0) & x \neq 0 \\ ([-1,1], 0) & x = 0 \end{cases}$$

and therefore

$$p_t + H(p_x, 0) > 0 \quad \text{for } q \in (-1, 1) \subset D^- u(0, t)$$

so that $|x|$ is not a viscosity solution to $\partial_t u + H(\partial_x u, x) = 0$.

8.1.6.1 The Pontryagin Principle for Generalized Solutions

Assume that X_* and α_* is an optimal control solution. Let

$$\begin{aligned} -\dot{\lambda}_*^t &= \partial_X f(X_*^t, \alpha_*^t)\lambda_*^t + \partial_X h(X_*^t, \alpha_*^t) \quad t < T, \\ \lambda_*^T &= g'(X_*^T). \end{aligned}$$

The proof of Theorem 8.13 shows first that $\Big(\lambda_*^t, -\big(\lambda_*^t \cdot f(X_*^t, \alpha_*^t) + h(X_*^t, \alpha_*^t)\big)\Big)$ is the first variation in x and t of the value function at the point (X^t, t) and concludes then that the first variation is in the superdifferential, that is

$$\Big(\lambda_*^t, -\big(\lambda_*^t \cdot f(X_*^t, \alpha_*^t) + h(X_*^t, \alpha_*^t)\big)\Big) \in \mathrm{D}^+ u(X_*^t, t).$$

Since the value function is a viscosity solution we conclude that

$$-\big(\lambda_*^t \cdot f(X_*^t, \alpha_*^t) + h(X_*^t, \alpha_*^t)\big) + \underbrace{H(\lambda_*^t, x)}_{\min_{\alpha \in A}\left(\lambda_*^t \cdot f(X_*^t, \alpha_*^t) + h(X_*^t, \alpha_*^t)\right)} \geq 0$$

which means that α_* satisfies the Pontryagin principle also in the case of non differentiable solutions to Hamilton-Jacobi equations.

8.1.6.2 Semiconcave Value Functions

There is an alternative and maybe more illustrative proof of the last theorem in a special setting: namely when the set of backward optimal paths $\{(\bar{X}^t, t) : t < T\}$, solving (8.29) and (8.47), may collide into a codimension one surface Γ in space-time $\mathbb{R}^d \times [0, T]$. Assume the value function is attained by precisely one path for $(x, t) \in \mathbb{R}^d \times [0, T] - \Gamma$ and that the minimum is attained by precisely two paths at $(x, t) \in \Gamma$. Colliding backward paths (or characteristics) X in general lead to a discontinuity in the gradient of the value function, $\lambda = u_x$, on the surface of collision, which means that the surface is a shock wave for the multidimensional system of conservation laws

$$\partial_t \lambda^i(x, t) + \frac{\mathrm{d}}{\mathrm{d}x_i} H\big(\lambda(x, t), x\big) = 0 \quad (x, t) \in \mathbb{R}^d \times [0, T], \ i = 1, \ldots, d.$$

Denote the jump, for fixed t, of a function w at Γ by $[w]$. To have two colliding paths at a point on Γ requires that λ has a jump $[\lambda] \neq 0$ there, since $[\lambda] = 0$ yields only one path. The implicit function theorem shows that for fixed t any compact subset of the set $\Gamma(t) \equiv \Gamma \cap (\mathbb{R}^d \times \{t\})$ is a $\mathcal{C}^1$ surface: the surface $\Gamma(t)$ is defined by the value functions, u^1

and u^2 for the two paths colliding on Γ, being equal on Γ and there are directions $\hat{n} \in \mathbb{R}^d$ so that the Jacobian determinant $\hat{n} \cdot \nabla(u^1 - u^2) = \hat{n} \cdot [\lambda] \neq 0$. Therefore compact subsets of the surface $\Gamma(t)$ has a well defined unit normal n. We assume that $\Gamma(t)$ has a normal everywhere and we will prove that $[\lambda] \cdot n \leq 0$, which implies that u is semi-concave.

Two optimal backwards paths that collide on $(x,t) \in \Gamma$ must depart in opposite direction away from Γ, that is $n \cdot H_\lambda(\lambda_+, x) \geq 0$ and $n \cdot H_\lambda(\lambda_-, x) \leq 0$, see Figure 8.6, so that

$$0 \leq n \cdot [H_\lambda(\lambda, x)] = n \cdot \underbrace{\int_0^1 H_{\lambda\lambda}(\lambda_- + s[\lambda])\, \mathrm{d}s [\lambda]}_{=:\bar{H}_{\lambda\lambda} \leq 0}. \tag{8.19}$$

We know that u is continuous also around Γ, therefore the jump of the gradient, $[u_x]$, has to be parallel to the normal, n, of the surface Γ. Lemma 8.28 shows that $[u_x] = [\lambda]$ and we conclude that this jump $[\lambda]$ is parallel to n so that $[\lambda] = [\lambda \cdot n]n$, which combined with (8.19) shows that

$$0 \leq [\lambda \cdot n] \bar{H}_{\lambda\lambda}\, n \cdot n.$$

The λ-concavity of the Hamiltonian, see Exercise 8.19, implies that the matrix $H_{\lambda\lambda}$ is negative semidefinite and consequently

$$\bar{H}_{\lambda\lambda}\, n \cdot n \leq 0, \tag{8.20}$$

which proves the claim $[\lambda] \cdot n \leq 0$, if we can exclude equality in (8.20). Equality in (8.20) means that $\bar{H}_{\lambda\lambda}\, n = 0$ and implies $H_\lambda(\lambda^+(t), x) = H_\lambda(\lambda^-(t), x)$ which is not compatible with two outgoing backward paths. Hence equality in (8.20) is ruled out. This derivation can be extended to several paths colliding into one point, see Exercise 8.18.

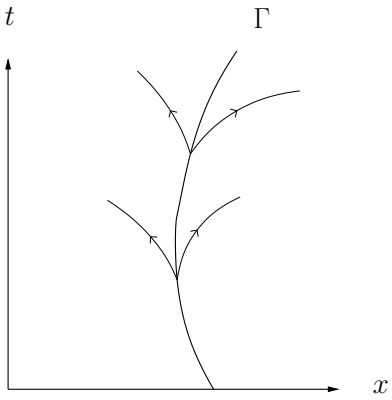


Figure 8.6: Optimal paths departing away from Γ.

Exercise 8.18.

Exercise 8.19. Show that the Hamiltonian

$$H(\lambda, x) := \min_{\alpha \in A} \big(\lambda \cdot f(x, \alpha) + h(x, \alpha)\big)$$

is concave in the λ-variable, that is show that for each λ^1 and λ^2 in $\mathbb{R}^d$ and for all $s \in [0, 1]$ there holds

$$H\big(s\lambda^1 + (1-s)\lambda^2, x\big) \geq sH(\lambda^1, x) + (1-s)H(\lambda^2, x).$$

8.1.7 Maximum Norm Stability of Viscosity Solutions

An important aspect of the viscosity solution of the Hamilton-Jacobi-Bellman equation is its maximum norm stability with respect to maximum norm perturbations of the data, in this case the Hamiltonian and the initial data; that is the value function is stable with respect to perturbations of the flux f and cost functions h and g.

Assume first for simplicity that the optimal control is attained and that the value function is differentiable for two different optimal control problems with data f, h, g and the Hamiltonian H, respectively $\bar{f}, \bar{h}, \bar{g}$ and Hamiltonian $\bar{H}$. The general case with only superdifferentiable value functions is studied afterwards. We have for the special case with the same initial data $\bar{X}^0 = X^0$ and $\bar{g} = g$

$$\begin{aligned}
&\underbrace{\int_0^T \bar{h}(\bar{X}^t, \bar{\alpha}^t)\,\mathrm{d}t + \bar{g}(\bar{X}^T)}_{\bar{u}(\bar{X}^0,0)} - \underbrace{\int_0^T h(X^t, \alpha^t)\,\mathrm{d}t + g(X^T)}_{u(X^0,0)} \\
&= \int_0^T \bar{h}(\bar{X}^t, \bar{\alpha}^t)\,\mathrm{d}t + u(\bar{X}^T, T) - \underbrace{u(X^0, 0)}_{u(\bar{X}^0,0)} \\
&= \int_0^T \bar{h}(\bar{X}^t, \bar{\alpha}^t)\,\mathrm{d}t + \int_0^T \mathrm{d}u(\bar{X}^t, t) \\
&= \int_0^T \underbrace{\partial_t u(\bar{X}^t, t)}_{=-H\left(\partial_x u(\bar{X}^t,t), \bar{X}^t\right)} + \underbrace{\partial_x u(\bar{X}^t, t) \cdot \bar{f}(\bar{X}^t, \bar{\alpha}^t) + \bar{h}(\bar{X}^t, \bar{\alpha}^t)}_{\geq \bar{H}\left(\partial_x u(\bar{X}^t,t), \bar{X}^t\right)}\,\mathrm{d}t \\
&\geq \int_0^T (\bar{H} - H)\big(\partial_x u(\bar{X}^t, t), \bar{X}^t\big)\,\mathrm{d}t.
\end{aligned} \tag{8.21}$$

The more general case with $\bar{g} \neq g$ yields the additional error term

$$(g - \bar{g})(\bar{X}^T)$$

to the right hand side in (8.21).

To find an upper bound, repeat the derivation above, replacing u along $\bar{X}^t$ with $\bar{u}$

along X^t, to obtain

$$\begin{aligned}
&\underbrace{\int_0^T h(X^t,\alpha^t)\,\mathrm{d}t + g(X^T)}_{u(X^0,0)} - \underbrace{\int_0^T \bar{h}(\bar{X}^t,\bar{\alpha}^t)\,\mathrm{d}t + \bar{g}(\bar{X}^T)}_{\bar{u}(\bar{X}^0.0)}\\
&= \int_0^T h(X^t,\alpha^t)\,\mathrm{d}t + \bar{u}(X^T,T) - \underbrace{\bar{u}(\bar{X}^0,0)}_{\bar{u}(X^0,0)}\\
&= \int_0^T h(X^t,\alpha^t)\,\mathrm{d}t + \int_0^T \mathrm{d}\bar{u}(X^t,t)\\
&= \int_0^T \underbrace{\partial_t \bar{u}(X^t,t)}_{=-\bar{H}\left(\partial_x\bar{u}(X^t,t),X^t\right)} + \underbrace{\partial_x\bar{u}(X^t,t)\cdot f(X^t,\alpha^t) + h(X^t,\alpha^t)}_{\geq H\left(\partial_x\bar{u}(X^t,t),X^t\right)}\,\mathrm{d}t\\
&\geq \int_0^T (H-\bar{H})\big(\partial_x\bar{u}(X^t,t),X^t\big)\,\mathrm{d}t.
\end{aligned}$$

The two estimates above yields both an upper and a lower bound

$$\begin{aligned}
\int_0^T (H-\bar{H})\big(\partial_x\bar{u}(X^t,t),X^t\big)\,\mathrm{d}t &\leq u(X_0,0) - \bar{u}(X_0,0)\\
&\leq \int_0^T (H-\bar{H})\big(\partial_x u(\bar{X}^t,t),\bar{X}^t\big)\,\mathrm{d}t.
\end{aligned} \tag{8.22}$$

Remark 8.20 (No minimizers)**.** If there are no minimizers (α, X) and $(\bar{\alpha},\bar{X})$, then for every $\varepsilon > 0$, we can choose controls $\alpha,\bar{\alpha}$ with corresponding states $X,\bar{X}$ such that

$$E_{lhs} - \varepsilon \leq u(X_0,0) - \bar{u}(X_0,0) \leq E_{rhs} + \varepsilon$$

with E_{lhs}, E_{rhs} being the left and right hand sides of (8.22).

Solutions to Hamilton-Jacobi equations are in general not differentiable as we have seen in Example 8.9. Let us extend the derivation of (8.22) to a case when u is not differentiable. If u is a non differentiable semiconcave solution to a Hamilton-Jacobi equation, Definition 8.12 of the viscosity solution reduces to

$$\begin{aligned}
&p_t + H(p_x,x) = 0 \quad \text{for all } (p_t,p_x)\in \mathrm{D}u(x,t) \text{ and all } t<T, x\in\mathbb{R}^d,\\
&p_t + H(p_x,x) \geq 0 \quad \text{for all } (p_t,p_x)\in \mathrm{D}^+u(x,t) \text{ and all } t<T, x\in\mathbb{R}^d,\\
&u(\cdot,T) = g.
\end{aligned}$$

Consider now a point (x,t) where the value function is not differentiable. This means that in (8.21) we can for each t choose a point $(p_t,p_x)\in \mathrm{D}^+u(X^t,t)$ so that

$$\begin{aligned}
\int_0^T \mathtt{d}u(\bar{X}^t,t) + \int_0^T \bar{h}(\bar{X}^t,\bar{\alpha}^t)\,\mathtt{d}t &= \int_0^T \big(p_t + p_x\cdot\bar{f}(\bar{X}^t,\bar{\alpha}^t) + \bar{h}(\bar{X}^t,\bar{\alpha}^t)\big)\,\mathtt{d}t\\
&\geq \int_0^T \big(p_t + \bar{H}(p_x,\bar{X}^t)\big)\,\mathtt{d}t \geq \int_0^T \big(-H+\bar{H}\big)(p_x,\bar{X}^t)\,\mathtt{d}t\,.
\end{aligned}$$

Note that the only difference compared to the differentiable case is the inequality instead of equality in the last step, which uses that optimal control problems have semi-concave viscosity solutions. The analogous formulation holds for $\bar{u}$. Consequently (8.22) holds for some $(p_t, p_x) \in \mathrm{D}^+u(\bar{X}^t, t)$ replacing $(\partial_t u(\bar{X}^t, t), \partial_x u(\bar{X}^t, t))$ and some $(\bar{p}_t, \bar{p}_x) \in \mathrm{D}^+\bar{u}(X^t, t)$ replacing $\big(\partial_t \bar{u}(X^t, t), \partial_x \bar{u}(X^t, t)\big)$.

The present analysis is in principle valid even when we replace $\mathbb{R}^d$ to be an infinite dimensional Hilbert space for optimal control of partial differential equations, although existence and semiconcavity of solutions is not derived in full generality, see [36]

8.2 Numerical Approximation of ODE Constrained Minimization

We consider numerical approximations with the time steps

$$t_n = \frac{n}{N}T, \quad n = 0, 1, 2, \dots, N.$$

The most basic approximation is based on the minimization

$$\min_{\bar{\alpha} \in B^N} \big(g(\bar{X}_N) + \sum_{n=0}^{N-1} h(\bar{X}_n, \bar{\alpha}_n)\Delta t\big), \tag{8.23}$$

where $\Delta t = t_{n+1} - t_n$, $\bar{X}_0 = X_0$ and $\bar{X}_n \equiv \bar{X}(t_n)$, for $1 \le n \le N$, satisfy the *forward Euler* constraint

$$\bar{X}_{n+1} = \bar{X}_n + \Delta t\, f(\bar{X}_n, \bar{\alpha}_n). \tag{8.24}$$

The existence of at least one minimum of (8.23) is clear since it is a minimization of a continuous function in the compact set B^N. The Lagrange principle can be used to solve such a constrained minimization problem. We will focus on a variant of this method based on the *discrete Pontryagin principle* where the control is eliminated

$$\begin{aligned} \bar{X}_{n+1} &= \bar{X}_n + \Delta t H_\lambda\big(\bar{\lambda}_{n+1}, \bar{X}_n\big), \quad \bar{X}_0 = X_0, \\ \bar{\lambda}_n &= \bar{\lambda}_{n+1} + \Delta t H_x\big(\bar{\lambda}_{n+1}, \bar{X}_n\big), \quad \bar{\lambda}_N = g_x(\bar{X}_N), \end{aligned} \tag{8.25}$$

called the symplectic Euler method for the Hamiltonian system (8.13), cf. [20].

A natural question is in what sense the discrete problem (8.25) is an approximation to the continuous optimal control problem (8.13). In this section we show that the value function of the discrete problem approximates the continuous value function, using the theory of viscosity solutions to Hamilton-Jacobi equations to construct and analyse regularized Hamiltonians.

Our analysis is a kind of backward error analysis. The standard backward error analysis for Hamiltonian systems uses an analytic Hamiltonian and shows that symplectic one step schemes generate approximate paths that solve a modified Hamiltonian system, with the perturbed Hamiltonian given by a series expansion cf. [20]. Our backward error analysis is different and more related to the standard finite element analysis. We first

extend the approximate Euler solution to a continuous piecewise linear function in time and define a *discrete value function*, $\bar{u} : \mathbb{R}^d \times [0,T] \to \mathbb{R}$. This value function satisfies a perturbed Hamilton-Jacobi partial differential equation, with a small residual error. A special case of our analysis shows that if the optimal α in (8.5) is a differentiable function of x and λ and if the optimal backward paths, $\bar{X}(s)$ for $s < T$, do not collide, more about this later, the discrete value functions, $\bar{u}$, for the Pontryagin method (8.25) satisfies a Hamilton-Jacobi equation:

$$\bar{u}_t + H(\bar{u}_x, \cdot) = \mathcal{O}(\Delta t), \qquad \text{as } \Delta t \to 0+, \tag{8.26}$$

where

$$\bar{u}(x, t_m) \equiv \min_{\bar{\alpha} \in B^N} \left(g(\bar{X}_N) + \sum_{n=m}^{N-1} h(\bar{X}_n, \bar{\alpha}_n) \Delta t \right) \tag{8.27}$$

for solutions $\bar{X}$ to with $\bar{X}(t_m) \equiv \bar{X}_m = x$. The minimum in (8.27) is taken over the solutions to the discrete Pontryagin principle (8.25). The maximum norm stability of Hamilton–Jacobi PDE solutions and a comparison between the two equations (8.4) and (8.26) show that

$$\mathcal{O} \left\| u - \bar{u} \right\|_{\mathcal{C}} = \mathcal{O}(\Delta t). \tag{8.28}$$

However, in general the optimal controls $\bar{\alpha}$ and α in (8.24) and (8.1) are discontinuous functions of x, and $\bar{\lambda}$ or u_x, respectively, and the backward paths *do* collide. There are two different reasons for discontinuous controls:

- The Hamiltonian is in general only Lipschitz continuous, even if f and h are smooth.
- The optimal backward paths may collide.

The standard error analysis for ordinary differential equations is directly applicable to control problems when the time derivative of the control function is integrable. But general control problems with discontinuous controls require alternative analysis, which will be in two steps. The *first step* in our error analysis is to construct regularizations of the functions f and h, based on (8.14) applied to a $\mathcal{C}^2(\mathbb{R}^d \times \mathbb{R}^d)$ approximate Hamiltonian H^δ which is λ-concave and satisfies

$$\|H^\delta - H\|_{\mathcal{C}} = \mathcal{O}(\delta), \quad \text{as } \delta \to 0^+,$$

and to introduce the regularized paths

$$\begin{aligned} \bar{X}_{n+1} &= \bar{X}_n + \Delta t H^\delta_\lambda\big(\bar{\lambda}_{n+1}, \bar{X}_n\big), \quad \bar{X}_0 = X_0, \\ \bar{\lambda}_n &= \bar{\lambda}_{n+1} + \Delta t H^\delta_x\big(\bar{\lambda}_{n+1}, \bar{X}_n\big), \quad \bar{\lambda}_N = g_x(\bar{X}_N). \end{aligned} \tag{8.29}$$

We will sometimes use the notation $f^\delta \equiv H^\delta_\lambda$ and $h^\delta \equiv H^\delta - \lambda H^\delta_\lambda$.

The *second step* is to estimate the residual of the discrete value function in the Hamilton-Jacobi-Bellman equation (8.4). The maximum norm stability of viscosity solutions and the residual estimate imply then an estimate for the error in the value

function. An approximation of the form (8.29) may be viewed as a general symplectic one step method for the Hamiltonian system (8.13), see Section 8.2.7.

There is a second reason to use Hamiltonians with smooth flux: in practice the nonlinear boundary value problem (8.29) has to be solved by iterations. If the flux is not continuous it seems difficult to construct a convergent iterative method, in any case iterations perform better with smoother solutions. When the Hamiltonian can be formed explicitly, the Pontryagin based method has the advantage that the Newton method can be applied to solve the discrete nonlinear Hamiltonian system with a sparse Jacobian.

If the optimal discrete backward paths $\bar{X}(t)$ in (8.29) collide on a codimension one surface Γ in $\mathbb{R}^d \times [0,T]$, the dual variable $\bar{\lambda} = \bar{u}_x$ may have a discontinuity at Γ, as a function of x. Theorems 8.27 and **??** prove, for $\bar{u}$ based on the Pontryagin method, that in the viscosity solution sense

$$\bar{u}_t + H(\bar{u}_x, \cdot) = \mathcal{O}(\Delta t + \delta + \frac{(\Delta t)^2}{\delta}), \tag{8.30}$$

where the discrete value function, $\bar{u}$, in (8.27) has been modified to

$$\bar{u}(x, t_m) = \min_{\bar{X}_m = x} \Big(g(\bar{X}_N) + \sum_{n=m}^{N-1} h^\delta(\bar{X}_n, \bar{\lambda}_{n+1}) \Delta t \Big). \tag{8.31}$$

The regularizations make the right hand side in (8.30) a Lipschitz continuous function of $\big(\bar{\lambda}(t), \bar{X}(t), t\big)$, bounded by $C(\Delta t + \delta + \frac{(\Delta t)^2}{\delta})$ where C depends only on the Lipschitz constants of f, h and $\bar{\lambda}$. Therefore the maximum norm stability can be used to prove $\|u - \bar{u}\|_{\mathcal{C}} = \mathcal{O}(\Delta t)$, for $\delta = \Delta t$. Without the regularization, the corresponding error term to in (8.30) is not well defined, even if $\bar{u}_x$ is smooth. A similar proof applies to the minimization method for smooth Hamiltonians, see [36]. It is important to note that for non smooth control the solution paths $\bar{X}$ may not converge although the value function converges as Δt and δ tend to zero. Therefore our backward error analysis uses consistency with the Hamilton-Jacobi partial differential equation and not with the Hamiltonian system. Convergence of the approximate path $(\bar{X}, \bar{\lambda})$ typically requires Lipschitz continuous flux (H_λ, H_x), which we do not assume in this work.

8.2.1 Optimization Examples

We give some examples when the Hamiltonian, H, is not a differentiable function, and difficulties associated with this.

Example 8.21. Let $B = \{-1, 1\}$, $f = \alpha$, $h = x^2/2$ and $g = 0$. Here the continuous minimization problem (8.3) has no minimizer among the measurable functions. A solution in discrete time using a nonregularized Pontryagin method or discrete dynamic programming will behave as in Figure 8.7. First the solution approaches the time axis, and then it oscillates back and forth. As Δt becomes smaller these oscillations do so as well. The infimum for the continuous problem corresponds to a solution $X(t)$ that approaches the time-axis, and then remains on it. However, this corresponds to $\alpha = 0$,

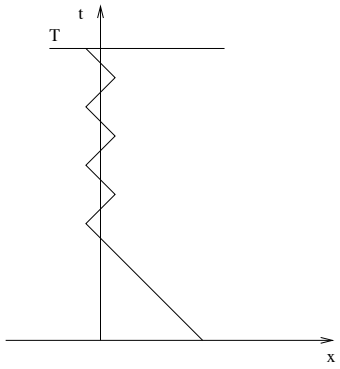


Figure 8.7: Example 8.21 where the continuous problem has no minimizer among the measurable functions.

which is not in B, and hence the infimum is not attained. A cure to always have an attained minimizing path for the continuous problem is to use controls which are Young measures, see [41] and [31]. We note that the Hamiltonian, $H(\lambda, x) = -|\lambda| + x^2/2$, in this example is not differentiable.

Example 8.22. Let $B = [-1, 1]$, $f = \alpha$, $h = x^2/2$ and $g = 0$, which is similar to the previous example but now the set of admissible controls, B, has been changed slightly. Since $0 \in B$, the infimum in (8.3) is now obtained. However, the Hamiltonian remains unchanged compared to the previous example, and a solution to the discrete Pontryagin principle would still be oscillating as in Figure 8.7.

Example 8.23. Let $B = [-1, 1]$, $f = \alpha$, $h = 0$ and $g = x^2$. The Hamiltonian is nondifferentiable: $H = -|\lambda|$. If $T = 1$ there are infinitely many solutions to the continuous minimization, the discrete minimization and the unregularized discrete Pontryagin principle, when $X_0 \in (-1, 1)$, as depicted in Figure 8.8.

The problems occurring in the previous examples are all cured by regularizing the Hamiltonian and using the scheme (8.29). That is, the solution to (8.29) in the first two examples is a smooth curve that obtains a increasingly sharp kink near the time-axis as the regularizing parameter, δ, decreases, see Figure 8.9. In the last of the previous examples we, in contrast to the other methods, obtain a unique solution to (8.29).

Another problem that has not to do with nondifferentiability of the Hamiltonian is shown in the following example:

Example 8.24. Let $B = [-1, 1]$, $f = \alpha$, $h = 0$ and $g = -|x|$. Although H is discontinuous here, this is not what causes problem. The problem is that optimal paths collide backwards, see Figure 8.10. When $X_0 = 0$ there are two solutions, one going to the left, and one to the right. The left solution has $\lambda = 1$ and the right solution has $\lambda = -1$, so on the time-axis λ is discontinuous. For these values of λ, the Hamiltonian is differentiable, therefore the nonsmoothness of the Hamiltonian is not the issue here. It is rather the global properties of the problem that play a role. This problem is difficult to

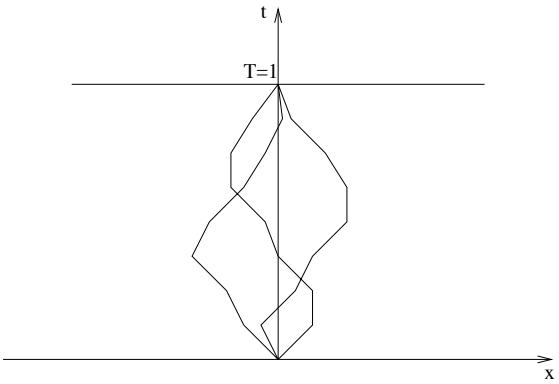


Figure 8.8: Example 8.23 with $g(x) = x^2$ gives infinitely many minimizing paths through the same starting point.

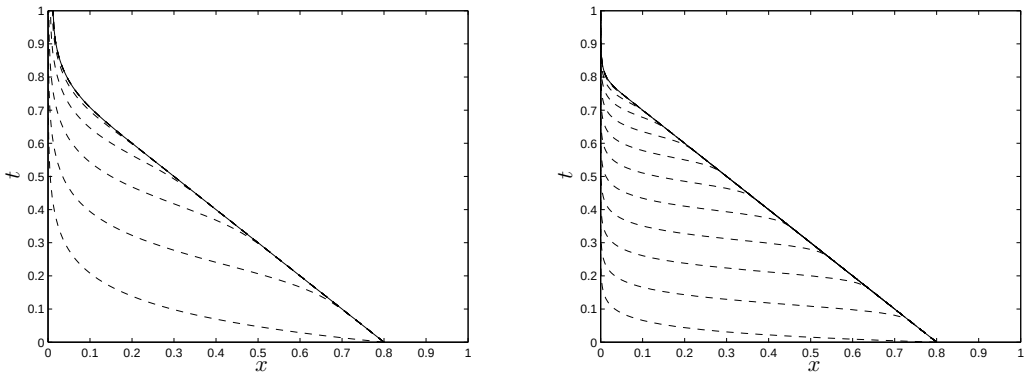


Figure 8.9: Solution of the discrete optimization problem (8.29) in Example 8.21 and 8.22 for $\delta = \Delta t = 1/N$, $X_0 = 0.8$ and $H_\lambda^\delta(\lambda, x) = -\tanh(\lambda/\delta)$, using the Newton method. To the left, $N = 100$, and to the right, $N = 1000$. The dashed lines shows the solution after each Newton iteration.

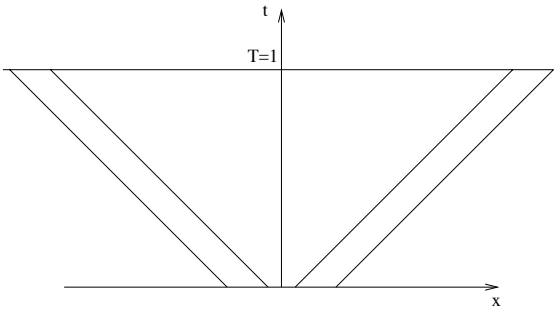


Figure 8.10: Solution of the optimization problem in Example 8.24, where $g(x) = -|x|$, $f = \alpha$, $h = 0$ and $B = [-1, 1]$, for four different starting points.

regularize, and it will not be done here. However, we still can show convergence of the scheme (8.29). This is done in Section **??**.

When using (8.29) to solve the minimization problem (8.3) it is assumed that the Hamiltonian is exactly known. Is this an unrealistic assumption in practice? In the following two examples we indicate that there exist interesting examples where we know the Hamiltonian. The first has to do with volatility estimation in finance, and the latter with optimization of an electric contact.

8.2.1.1 Implied Volatility

Black-Scholes equation for pricing general options uses the volatility of the underlying asset. This parameter, however, is difficult to estimate. One way of estimation is to use measured market values of options on the considered asset for standard European contracts. This way of implicitly determining the volatility is called implied volatility. In the simplest setting, the formula[2] for the option price based on constant interest rate and volatility is used. Then the result typically gives different values of the volatility for different stock price – instead of obtaining a constant volatility, the implied volatility becomes a strictly convex function of the stock price called the volatility smile. Below we shall fit a model allowing the volatility to be a general function to observed option prices. That requires solution of a partial differential equation, since an explicit formula is not available. Another ingredient in our reconstruction is to use the so called Dupire equation for standard European put and call option prices as a function of the strike price and strike time. Using an equation of the option value as a function of the strike price and strike time, for given stock price, is computational more efficient, since the option data is for different strike price and strike times, with fixed stock price. To use the standard Black-Scholes equation for the option value as a function of the stock price

[2] the option price formula is $C(s,t;K,T) = s\Phi(d_1) - Ke^{-r(T-t)}\Phi(d_2)$, where $d_1 := \big(\ln(s/K) + (r + \sigma^2/2)(T-t)\big)/\big(\sigma(T-t)^{1/2}\big)$, $d_2 := d_1 - \sigma(T-t)^{1/2}$ and Φ is the standard normal cumulative distribution function.

would require to solve different equations for each data point, which is also possible but more computationally expensive.

We assume that the financial asset obeys the following Ito stochastic differential equation,

$$dS(t) = \mu S(t)dt + \sigma\big(t, S(t)\big)S(t)dW(t), \tag{8.32}$$

where $S(t)$ is the price of the asset at time t, μ is a drift term, σ is the volatility and $W : \mathbb{R}_+ \to \mathbb{R}$ is the Wiener process. If the volatility is a sufficiently regular function of S, t, the strike level K and the maturity date T, the Dupire equation holds for the option price $C(T, K)$ as a function of T and K, with the present time $t = 0$ and stock price $S(0) = S$ fixed,

$$\begin{aligned} &C_T - \tilde{\sigma} C_{KK} = 0, \quad T \in (0, \infty), K > 0, \\ &C(0, K) = \max\{S - K, 0\} \quad K > 0, \end{aligned} \tag{8.33}$$

where

$$\tilde{\sigma}(T, K) \equiv \frac{\sigma^2(T, K)K^2}{2}.$$

Here the contract is an european call option with payoff function $\max\{S(T) - K, 0\}$. We have for simplicity assumed the bank rate to be zero. A derivation of Dupire's equation (8.33) is presented in Example 8.25 in the special setting $r = 0$; the general case is studied in [10].

The optimization problem now consists of finding $\sigma(T, K)$ such that

$$\int_0^{\hat{T}} \int_{\mathbb{R}_+} (C - \hat{C})^2(T, K)w(T, K)dKdT \tag{8.34}$$

is minimized, where $\hat{C}$ are the measured market values on option prices for different strike prices and strike times and w is a non negative weight function. In practice, $\hat{C}$ is not known everywhere, but for the sake of simplicity, we assume it is and set $w \equiv 1$, that is there exists a future time $\hat{T}$ such that $\hat{C}$ is defined in $\mathbb{R}_+ \times [0, \hat{T}]$. If the geometric Brownian motion would be a perfect model for the evolution of the price of the asset, the function $\sigma(T, K)$ would be constant, but as this is not the case, the σ that minimizes (8.34) (if a minimizer exists) varies with T and K.

It is possible to use (8.13) and (8.25) to perform the minimization of (8.34) over the solutions to a finite difference discretization of (8.33)

$$\begin{aligned} &\min_{\tilde{\sigma}} \int_0^{\hat{T}} \Delta K \sum_i (C - \hat{C})_i^2 dT \\ &\text{subject to} \quad \frac{\partial C_i(T)}{\partial T} = \tilde{\sigma} D^2 C_i(T), \\ &\qquad\qquad\quad C_i(0) = \max(S - i\Delta K, 0), \end{aligned} \tag{8.35}$$

where we now let $C_i(T) \approx C(T, i\Delta K)$ denote the discretized prize function, for strike time T and strike price $i\Delta K$, and D^2 is the standard three point difference approximation

of the second order partial derivative in K, that is $(D^2C)_i = (C_{i+1} - 2C_i + C_{i-1})/\Delta K^2$. In order to have a finite dimensional problem we restrict to a compact interval $(0, M\Delta K)$ in K with the boundary conditions

$$C_0 = S, \quad C_M = 0.$$

This formulation will be exactly the same as in (8.13) if $\Delta K = 1$, and otherwise it requires to use a new scalar product $(x, y) := \Delta K \sum_i x_i y_i$ and let the partial derivative ∂_λ be replaced by the following Gateaux derivative, H_λ,

$$\lim_{\epsilon \to 0} \epsilon^{-1} \big(H(\lambda + \epsilon v, C) - H(\lambda, C) \big) =: \big(H_\lambda(\lambda, C), v \big),$$

and similarly for ∂_C; so that the partial derivative is a factor of ΔK smaller than the Gateaux derivative. This complication with using $\Delta K \neq 1$ is introduced in order to have a consistent formulation with the infinite dimensional case, where a partial derivative of a functional becomes zero but the Gateaux derivative is nonzero and meaningful, see the next example. The reader may avoid this be considering $\Delta K = 1$.

The Hamiltonian for this problem is

$$\begin{aligned} H(\lambda, C) &= \Delta K \min_{\tilde{\sigma}} \sum_{i=1}^{M-1} \Big(\lambda_i \tilde{\sigma}_i (D^2 C)_i + (C - \hat{C})_i^2 \Big) \\ &= \Delta K \sum_{i=1}^{M-1} \Big(\min_{\tilde{\sigma}_i} \lambda_i \tilde{\sigma}_i (D^2 C)_i + (C - \hat{C})_i^2 \Big) \end{aligned}$$

where λ is the adjoint associated to the constraint (8.35). We have used that the components of the flux, f, in this problem is $\tilde{\sigma}_i (D^2 C)_i$, that the running cost, h, is $\Delta K \sum_i (C - \hat{C})_i^2$, and further that each $\tilde{\sigma}_i$ minimizes $\lambda_i \tilde{\sigma}_i (D^2 C)_i$ separately, so that the minimum can be moved inside the sum. If we make the simplifying assumption that $0 \leq \sigma_- \leq \tilde{\sigma} \leq \sigma_+ < \infty$ we may introduce a function $s : \mathbb{R} \to \mathbb{R}$ as

$$s(y) \equiv \min_{\tilde{\sigma}} y\, \tilde{\sigma} = \begin{cases} y\sigma_-, & y > 0 \\ y\sigma_+, & y < 0. \end{cases}$$

Using s, it is possible to write the Hamiltonian as

$$H(\lambda, C) = \Delta K \sum_{i=1}^{M-1} \Big(s\big(\lambda_i (D^2 C)_i\big) + \big(C - \hat{C}\big)_i^2 \Big).$$

Since s is nondifferentiable, so is H. However, s may easily be regularized, and it is possible to obtain the regularization in closed form, e.g. as in Example 1. Using a regularized version s_δ of s, the regularized Hamiltonian becomes

$$H^\delta(\lambda, C) = \Delta K \sum_{i=1}^{M-1} \Big(s_\delta\big(\lambda_i (D^2 C)_i\big) + \big(C - \hat{C}\big)_i^2 \Big),$$

which using Gateaux derivatives gives the Hamiltonian system

$$\begin{aligned}\frac{\partial C_i(T)}{\partial T} &= s'_\delta\big(\lambda_i (D^2 C)_i\big) D^2 C_i(T), \quad C_0 = S \quad C_M = 0,\\ -\frac{\partial \lambda_i(T)}{\partial T} &= D^2\Big(s'_\delta\big(\lambda_i (D^2 C)_i\big)\lambda\Big) + 2(C-\hat{C})_i,\\ \lambda_0 &= \lambda_M = 0,\end{aligned} \tag{8.36}$$

with data

$$C_i(0) = \max(S - i\Delta K, 0), \qquad \lambda(\hat{T}) = 0.$$

The corresponding Hamilton-Jacobi equation for the value function

$$u(C,\tau) = \int_\tau^{\hat{T}} \sum_{i=1}^{M-1} (C-\hat{C})_i^2 \Delta K dT$$

is

$$\begin{aligned} &u_T + H(u_C, C) = 0, \quad T < \hat{T},\\ &u(\hat{T}, \cdot) = 0,\end{aligned}$$

where u_C is the Gateaux derivative with respect to C in the scalar product $(x,y) \equiv \Delta K \sum_i x_i, y_i$. With this scalar product the Hamiltonian system (8.36) takes the form

$$\begin{aligned}(C_T, v) &= (H^\delta_\lambda, v), \quad \forall v \in \mathbb{R}^{M-1}\\ (\lambda_T, v) &= -(H^\delta_C, v), \quad \forall v \in \mathbb{R}^{M-1}\end{aligned}$$

where H^δ_λ and H^δ_C are the Gateaux derivatives.

A choice of the regularization parameter δ, depending also on data error, can be obtained e.g. by the discrepancy principle, cf. [39], [12]. The Newton method described in Section 3 works well to solve the discrete equations for $d = 10$. The results of one trial volatility estimation is given in Figure 8.11.

Example 8.25 (Derivation of Dupire's equation)**.** The Black-Scholes equation for a general volatility function and interest $r = 0$ is

$$\begin{aligned}\partial_t f + \frac{\sigma^2(s,t)s^2}{2}\partial_{ss} f = 0 \quad t < T\\ f(s,T) = \max(K-s, 0)\end{aligned}$$

which defines the option value $f(s,t;K,T)$. The goal is now to find the equation for f as a function of K and T. We know from the Kolmogorov backward equation that $f(s,t;K,T) = \mathbb{E}[\max(K - S_T, 0) \mid S_t = s]$, where $dS_t = \sigma(S_t,t)S_t dW_t$. The Kolmogorov forward equation shows that $f(s,t;K,T) = \int_{\mathbb{R}} \max(K-y,0)p(y,T;s,t)dy$ where

$$\begin{aligned}\partial_T p - \partial_{yy}\Big(\frac{\sigma^2(y,T)y^2}{2}p\Big) = 0 \quad T > t\\ p(y,t;s,t) = \delta(y-s).\end{aligned}$$

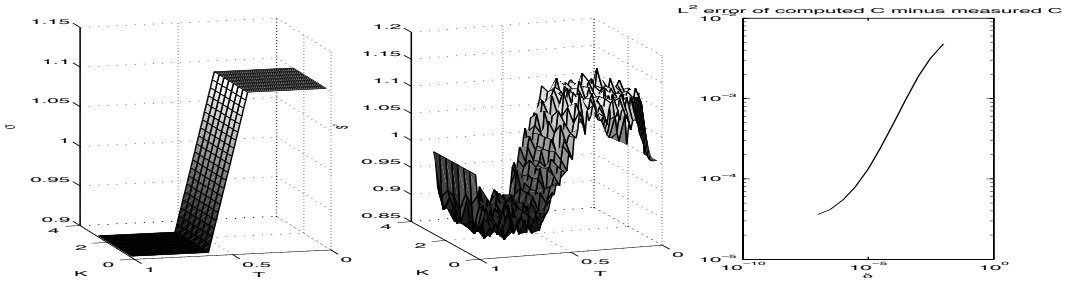


Figure 8.11: Results of a computer experiment where the volatility σ in the picture to the left is used to obtain the "measured" $\hat{C}$. Uniform noise of amplitude 10^{-4} is also added to $\hat{C}$. The error $\|C - \hat{C}\|_{L^2}$ is plotted versus δ in the picture to the right. In the middle picture the approximate volatility, s'_δ is shown for the value of δ $(= 3 \cdot 10^{-6})$ that minimizes $\|s'_\delta - \sigma\|_{L^2}$. In this experiment, $M = 9$ and $N = 100$.

We observe that $\partial_{KK} f(s,t;K,T) = \int_{\mathbb{R}} \delta(K-y) p(y,T;s,t) dy = p(K,T;s,t)$ and consequently

$$\partial_T \partial_{KK} f(s,t;K,T) - \partial_{KK}\Big(\frac{\sigma^2(K,T)K^2}{2}\partial_{KK} f(s,t;K,T)\Big) = 0 \quad T > t,$$

can be integrated to obtain

$$\partial_T f(s,t;K,T) - \Big(\frac{\sigma^2(K,T)K^2}{2}\partial_{KK} f(s,t;K,T)\Big) = C_1 + C_2 K \quad T > t.$$

The boundary condition $\partial_{KK} f \to 0$ as $K \to \infty$ and $\partial_T f \to 0$ as $T \to \infty$ concludes that $C_1 = C_2 = 0$.

8.2.1.2 Topology Optimization of Electric Conduction

The problem is to place a given amount of conducting material in a given domain $\Omega \subset \mathbb{R}^d$ in order to minimize the power loss for a given surface current q, satisfying $\int_{\partial\Omega} q ds = 0$: let $\eta \in \mathbb{R}$ be a given constant, associated to the given amount of material, and find an optimal conduction distribution $\sigma : \Omega \to \{\sigma_-, \sigma_+\}$, where $\sigma_\pm > 0$, such that

$$\begin{aligned} &\operatorname{div}\big(\sigma \nabla \varphi(x)\big) = 0, \ x \in \Omega, \qquad \sigma \frac{\partial \varphi}{\partial n}\Big|_{\partial\Omega} = q \\ &\min_\sigma \Big(\int_{\partial\Omega} q\varphi \, \mathrm{d}s + \eta \int_\Omega \sigma \, \mathrm{d}x\Big), \end{aligned} \tag{8.37}$$

where $\partial/\partial n$ denotes the normal derivative and $\mathrm{d}s$ is the surface measure on $\partial\Omega$. Note that (8.37) implies that the power loss satisfies

$$\begin{aligned} \int_{\partial\Omega} q\varphi \, \mathrm{d}s &= -\int_\Omega \operatorname{div}(\sigma\nabla\varphi)\varphi \, \mathrm{d}x + \int_{\partial\Omega} \sigma \frac{\partial\varphi}{\partial n}\varphi \, \mathrm{d}s \\ &= \int_\Omega \sigma \nabla\varphi \cdot \nabla\varphi \, \mathrm{d}x. \end{aligned}$$

The Lagrangian takes the form

$$\int_{\partial\Omega} q(\varphi+\lambda)\,\mathrm{d}s + \int_\Omega \sigma \underbrace{(\eta - \nabla\varphi\cdot\nabla\lambda)}_{v}\,\mathrm{d}x$$

and the Hamiltonian becomes

$$H(\lambda,\varphi) = \min_\sigma \int_\Omega \sigma v\,\mathrm{d}x + \int_{\partial\Omega} q(\varphi+\lambda)\,\mathrm{d}s = \int_\Omega \underbrace{\min_\sigma \sigma v}_{s(v)}\,\mathrm{d}x + \int_{\partial\Omega} q(\varphi+\lambda)\,\mathrm{d}s$$

with the regularization

$$H^\delta(\lambda,\varphi) = \int_\Omega s_\delta(\eta - \nabla\varphi\cdot\nabla\lambda)\,\mathrm{d}x + \int_{\partial\Omega} q(\varphi+\lambda)\,\mathrm{d}s,$$

depending on the concave regularization $s_\delta \in \mathcal{C}^2(\mathbb{R})$ as in Section 8.2.1.1. The value function

$$u(\varphi,\tau) = \int_\tau^T (\int_{\partial\Omega} q\varphi\,\mathrm{d}s + \eta\int_\Omega \sigma\,\mathrm{d}x)\,\mathrm{d}t$$

for the parabolic variant of (8.37), that is

$$\varphi_t = \mathrm{div}\big(\sigma\nabla\varphi(x)\big),$$

yields the infinite dimensional Hamilton-Jacobi equation

$$\partial_t u + H(\partial_\varphi u,\varphi) = 0 \quad t<T, \quad u(\cdot,T) = 0,$$

using the Gateaux derivative $\partial_\varphi u = \lambda$ of the functional $u(\varphi,t)$ in $L^2(\Omega)$. The regularized Hamiltonian generates the following parabolic Hamiltonian system for φ and λ

$$\begin{aligned}
\int_\Omega \big(\partial_t\varphi w + s'(\eta-\nabla\varphi\cdot\nabla\lambda)\nabla\varphi\cdot\nabla w\big)\,\mathrm{d}x &= \int_{\partial\Omega} qw\,\mathrm{d}s\\
\int_\Omega \big(-\partial_t\lambda v + s'(\eta-\nabla\varphi\cdot\nabla\lambda)\nabla\lambda\cdot\nabla v\big)\,\mathrm{d}x &= \int_{\partial\Omega} qv\,\mathrm{d}s
\end{aligned}$$

for all test functions $v,w\in V\equiv\{v\in H^1(\Omega)\;\big|\;\int_\Omega v dx = 0\}$. Time independent solutions satisfy $\lambda=\varphi$ by symmetry. Therefore the electric potential satisfies the nonlinear elliptic partial differential equation

$$\mathrm{div}\big(s'_\delta(\eta-|\nabla\varphi|^2)\nabla\varphi(x)\big) = 0 \quad x\in\Omega, \quad s'_\delta\frac{\partial\varphi}{\partial n}|_{\partial\Omega} = q, \tag{8.38}$$

which can be formulated as the convex minimization problem: $\varphi\in V$ is the unique minimizer (up to a constant) of

$$-\Big(\int_\Omega s_\delta(\eta-|\nabla\varphi(x)|^2)\,\mathrm{d}x + 2\int_{\partial\Omega} q\varphi\,\mathrm{d}s\Big). \tag{8.39}$$

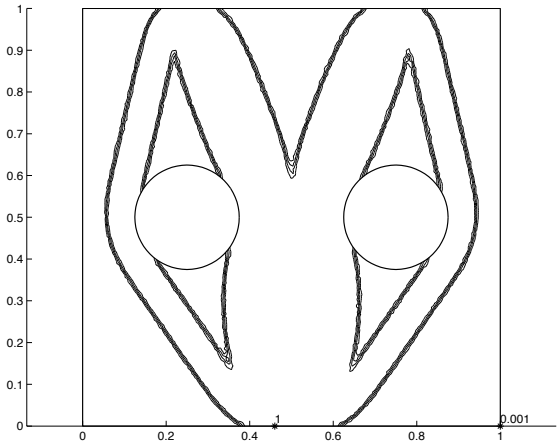


Figure 8.12: Contour plot of s'_δ as an approximation of the conductivity σ. As seen, Ω is in this example a square with two circles cut out. Electrical current enters Ω at two positions on the top of the square and leaves at one position on the bottom. The contours represent the levels 0.2, 0.4, 0.6 and 0.8. A piecewise linear FEM was used with 31440 elements, maximum element diameter 0.01, $\sigma_- = 0.001$, $\sigma_+ = 1$, $\eta = 0.15$ and $\delta = 10^{-5}$.

In [7] we study convergence of

$$\lim_{T\to\infty} \frac{u(\varphi,t) - \bar{u}(\varphi,t)}{T},$$

where $\bar{u}$ is the value function associated to finite element approximations of the minimization (8.39).

The Newton method in Section 3 works well to solve the finite element version of (8.38) by successively decreasing δ, also for large d, see [7], where also the corresponding inverse problem to use measured approximations of φ to determine the domain where $\sigma = \sigma_-$ and $\sigma = \sigma_+$ is studied. A numerical solution to (8.38) can be seen in Figure 8.12.

In this paper we use the standard Euclidean norm in $\mathbb{R}^d$ to measure X and λ. Optimal control of partial differential equations with X and λ belonging to infinite dimensional function spaces requires a choise of an appropriate norm. In [36] the analysis here is extended to optimal control of some parabolic partial differential equations, by replacing the Euclidean $\mathbb{R}^d$ norm with the H_0^1 Sobolev norm, using also that the theory of viscosity solutions remains valid with this replacement.

8.2.2 Solution of the Discrete Problem

We assume in the theorems that the Pontryagin minimization (8.29) has been solved exactly. In practice (8.29) can only be solved approximately by iterations. The simplest iteration method to solve the boundary value problem (8.29) is the shooting method:

start with an initial guess of $\bar{\lambda}[0]$ and compute, for all time steps n, the iterates

$$\begin{aligned} \bar{X}_{n+1} &= \bar{X}_n + \Delta t H^\delta_\lambda\big(\bar{\lambda}_{n+1}[i], \bar{X}_n\big), \quad n = 0, \dots, N-1, \quad \bar{X}_0 = X_0 \\ \bar{\lambda}_n[i+1] &= \bar{\lambda}_{n+1}[i] + \Delta t H^\delta_x\big(\bar{\lambda}_{n+1}[i], \bar{X}_n\big), \ n = N-1, \dots, 0, \ \bar{\lambda}_N = g_x(\bar{X}_N). \end{aligned} \tag{8.40}$$

An alternative method, better suited for many boundary value problems, is to use Newton iterations for the nonlinear system $F(\bar{X}, \bar{\lambda}) = 0$ where $F : \mathbb{R}^{Nd} \times \mathbb{R}^{Nd} \to \mathbb{R}^{2Nd}$ and

$$\begin{aligned} F(\bar{X}, \bar{\lambda})_{2n} &= \bar{X}_{n+1} - \bar{X}_n - \Delta t H^\delta_\lambda\big(\bar{\lambda}_{n+1}, \bar{X}_n\big), \\ F(\bar{X}, \bar{\lambda})_{2n+1} &= \bar{\lambda}_n - \bar{\lambda}_{n+1} - \Delta t H^\delta_x\big(\bar{\lambda}_{n+1}, \bar{X}_n\big). \end{aligned} \tag{8.41}$$

An advantage with the Pontryagin based method (8.41) is that the Jacobian of F can be calculated explicitly and it is sparse. The Newton method can be used to solve the volatility and topology optimization examples in Section 2, where the parameter δ is successively decreasing as the nonlinear equation (8.41) is solved more accurately.

Let us use dynamic programming to show that the system (8.29) has a solution in the case that $\bar{\lambda}$ is a Lipschitz continuous function of (x, t), with Lipschitz norm independent of Δt, and $\delta > C\Delta t$. One step

$$x = y + \Delta t H^\delta_\lambda\big(\lambda(x), y\big) \tag{8.42}$$

for fixed $y \in \mathbb{R}^d$ has a solution $x(y)$ since the iterations

$$x[i+1] = y + \Delta t H^\delta_\lambda\big(\lambda(x[i]), y\big)$$

yield a contraction for the error $e[i] = x[i+m] - x[i]$

$$e[i+1] = \Delta t\Big(H^\delta_\lambda\big(\lambda(x[i+m]), y\big) - H^\delta_\lambda\big(\lambda(x[i]), y\big)\Big) = \Delta t\overline{H^\delta_{\lambda\lambda}\lambda_x}e[i].$$

Conversely, for all $x \in \mathbb{R}^d$ equation (8.42) has a solution $y(x)$ for each step since the iterations

$$y[i+1] = x - \Delta t H^\delta_\lambda\big(\lambda(x), y[i]\big)$$

generate a contraction for the error. The dynamic programming principle then shows that there are unique paths through all points $\bar{X}_{n+1}$ leading to all $\bar{X}_n$ for all n.

Example 8.26. In Example 8.21 and 8.22 the problem was to minimize

$$\min_{\alpha \in B} \int_0^T \frac{X(t)^2}{2} dt,$$

given the dynamics

$$X'(t) = \alpha, \quad X(0) = X_0,$$

and an admissible set of controls $B = \{-1, 1\}$ (for Example 8.21), or $B = [-1, 1]$ (for Example 8.22). The Hamiltonian for this problem is $H(\lambda, x) = -|\lambda| + x^2/2$, and for a

smooth approximation of the λ-derivative, e.g. $H^\delta_\lambda(\lambda, x) = -\tanh(\lambda/\delta)$, the non-linear system (8.41) becomes

$$\begin{aligned} 0 &= \bar{X}_{n+1} - \bar{X}_n + \Delta t \tanh\left(\bar{\lambda}_{n+1}/\delta\right), \\ 0 &= \bar{\lambda}_n - \bar{\lambda}_{n+1} - \Delta t \bar{X}_n. \end{aligned}$$

Newton's method starts with an initial guess $\left(\bar{X}^0_{n+1}, \bar{\lambda}^0_n\right)$, for all times $n = 0, \ldots, N-1$, and updates the solution, for some damping factor $\gamma \in (0, 1]$, according to

$$\begin{aligned} \bar{X}^{i+1}_{n+1} &= \bar{X}^i_{n+1} - \gamma \Delta \bar{X}^i_{n+1}, \\ \bar{\lambda}^{i+1}_n &= \bar{\lambda}^i_n - \gamma \Delta \bar{\lambda}^i_n, \end{aligned}$$

where the updates comes from solving the sparse Newton system ($N = 3$ for illustration)

$$\begin{pmatrix} 1 & -1 & & & & \\ & d^i_1 \Delta t & 1 & & & \\ & 1 & -\Delta t & -1 & & \\ & & -1 & d^i_2 \Delta t & 1 & \\ & & & 1 & -\Delta t & \\ & & & & -1 & 1 \end{pmatrix} \begin{pmatrix} \Delta \bar{\lambda}^i_0 \\ \Delta \bar{\lambda}^i_1 \\ \Delta \bar{X}^i_1 \\ \Delta \bar{\lambda}^i_2 \\ \Delta \bar{X}^i_2 \\ \Delta \bar{X}^i_3 \end{pmatrix} = \begin{pmatrix} \bar{\lambda}^i_0 - \bar{\lambda}^i_1 - \Delta t \bar{X}^i_0 \\ \bar{X}^i_1 - \bar{X}^i_0 + \Delta t \tanh\left(\bar{\lambda}^i_1/\delta\right) \\ \bar{\lambda}^i_1 - \bar{\lambda}^i_2 - \Delta t \bar{X}^i_1 \\ \bar{X}^i_2 - \bar{X}^i_1 + \Delta t \tanh\left(\bar{\lambda}^i_2/\delta\right) \\ \bar{\lambda}^i_2 - \bar{\lambda}^i_3 - \Delta t \bar{X}^i_2 \\ \bar{X}^i_3 - \bar{X}^i_2 + \Delta t \tanh\left(\bar{\lambda}^i_3/\delta\right) \end{pmatrix},$$

and $d^i_j := \partial_\lambda \tanh\left(\bar{\lambda}^i_j/\delta\right) = \delta^{-1} \cosh^{-2}\left(\bar{\lambda}^i_j/\delta\right)$. A Matlab implementation for the above Newton method is shown below, and in Figure 8.9 the solution is shown for different values of N.

```
% Solving Hamiltonian system with Newton's method
% for T=1, delta=dt and gamma=1

N=1000; dt=1/N;
J=sparse(2*N,2*N);  rhs=sparse(2*N,1);
X=sparse(N+1,1); L=sparse(N+1,1);
X(1)= 0.8; % initial data

tol=1;
while tol>1e-6
  % Assemble Newton system row-wise
  for n=1:N
    rhs(2*n-1)=L(n)-L(n+1)-dt*X(n);
    rhs(2*n)=X(n+1)-X(n)+dt*tanh(L(n+1)/dt);
  end
  J(1,1:2)=[1 -1]; J(2*N,2*N-1:2*N)=[-1 1];
  for n=1:N-1
    J(2*n,2*n-1:2*n+1)=[-1 1/cosh(L(n+1)/dt)^2 1];
    J(2*n+1,2*n:2*n+2)=[1 -dt -1];
  end
```

```
  J(2,1)=0; J(2*N-1,2*N)=0;
  % Solve and update
  dXL=J\rhs;
  L(1)=L(1)-dXL(1); X(N+1)=X(N+1)-dXL(2*N);
  for n=2:N
    X(n)=X(n)-dXL(2*n-1); L(n)=L(n)-dXL(2*n-2);
  end
  tol = norm(rhs) % Error
end
```

8.2.3 Convergence of Euler Pontryagin Approximations

Theorem 8.27. *Assume that the Hamiltonian H, defined in (8.4), is Lipschitz continuous on $\mathbb{R}^d \times \mathbb{R}^d$ and that (8.29) has a solution $(\bar{X}, \bar{\lambda})$, where $\bar{\lambda}_{n+1}$ has uniformly bounded first variation with respect to $\bar{X}_n$ for all n and all Δt, that is there is a constant K such that*

$$|\partial_{\bar{X}_n} \bar{\lambda}_{n+1}| \leq K. \tag{8.43}$$

Then the optimal solution, $(\bar{X}, \bar{\lambda})$, of the Pontryagin method (8.29) satisfies the error estimate

$$\begin{aligned} \Big| \inf_{\alpha \in \mathcal{A}} \Big(g\big(X(T)\big) + \int_0^T h\big(X(s), \alpha(s)\big)\, \mathrm{d}s \Big) - \Big(g(\bar{X}_N) + \Delta t \sum_{n=0}^{N-1} h^\delta(\bar{X}_n, \bar{\lambda}_{n+1}) \Big) \Big| \\ = \mathcal{O}(\Delta t + \delta + \frac{(\Delta t)^2}{\delta}) \\ = \mathcal{O}(\Delta t), \quad \text{for } \delta = \Delta t. \end{aligned} \tag{8.44}$$

The bound $\mathcal{O}(\Delta t)$ in (8.44) depends on the dimension d through the Lipschitz norms of the Hamiltonian H and the constant K in (8.43).

The work [35] presents a convergence result for the case when backward paths $\bar{X}(t)$ collide on a $\mathcal{C}^1$ codimension one surface in $\mathbb{R}^d \times [0, T]$. The next subsections give a construction of a regularization H^δ and the proof of Theorem 8.27.

8.2.3.1 Construction of a Regularization

A possible regularization of H is to let H^δ be a standard convolution mollification of H

$$H^\delta(\lambda, x) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} H(z, y) \omega^\delta(z - \lambda) \omega^\delta(y - x)\, \mathrm{d}z\, \mathrm{d}y, \tag{8.45}$$

with $\omega^\delta : \mathbb{R}^d \to \mathbb{R}_+$ a $\mathcal{C}^2$ function compactly supported in the ball $\{y \in \mathbb{R}^d : |y| \leq \delta\}$ and with integral one $\int_{\mathbb{R}^d} \omega^\delta(y) dy = 1$. This regularization remains concave in λ. Our analysis is not dependent of this specific regularization, but uses that

$$\|H - H^\delta\|_C + \delta \|H^\delta\|_{C^1} + \delta^2 \|H^\delta\|_{C^2} = \mathcal{O}(\delta),$$

and that H^δ remains a concave function of λ.

8.2.3.2 Convergence without Shocks and Colliding Paths

The proof of the theorem is based on four lemmas. In all of those we suppose that the assumptions of Theorem 8.27 are valid.

Lemma 8.28. *The discrete dual function is the gradient of the value function, that is*

$$\bar{u}_x(\bar{X}_n, \bar{t}_n) = \bar{\lambda}_n. \tag{8.46}$$

Proof. The relation (8.46) holds for $t_n = T$. Use the induction assumption that (8.46) holds true for
$t_N \equiv T$, $t_{N-1}, \ldots, t_{n+1}$. Then the definitions of f^δ and h^δ imply

$$\begin{aligned}
\frac{\partial \bar{u}}{\partial \bar{X}_n}(\bar{X}_n, t_n) &= \partial_{\bar{X}_n}\big(\bar{u}(\bar{X}_{n+1}, t_{n+1}) + \Delta t h^\delta(\bar{\lambda}_{n+1}, \bar{X}_n)\big) \\
&= \partial_{\bar{X}_n}\bar{X}_{n+1}\frac{\partial \bar{u}}{\partial \bar{X}_{n+1}}(\bar{X}_{n+1}, t_{n+1}) + \Delta t \partial_{\bar{X}_n} h^\delta(\bar{\lambda}_{n+1}, \bar{X}_n) \\
&= \big(I + \Delta t \partial_{\bar{X}_n} H^\delta_\lambda(\bar{\lambda}_{n+1}, \bar{X}_n)\big)\bar{\lambda}_{n+1} + \Delta t \partial_{\bar{X}_n} h^\delta(\bar{\lambda}_{n+1}, \bar{X}_n) \\
&= \bar{\lambda}_{n+1} + \Delta t \partial_{\bar{X}_n}(H^\delta_\lambda \lambda + h^\delta)(\bar{\lambda}_{n+1}, \bar{X}_n) \\
&\quad - \Delta t H^\delta_\lambda(\bar{\lambda}_{n+1}, \bar{X}_n)\partial_{\bar{X}_n}\bar{\lambda}_{n+1} \\
&= \bar{\lambda}_{n+1} + \Delta t H^\delta_x(\bar{\lambda}_{n+1}, \bar{X}_n) \\
&= \bar{\lambda}_n.
\end{aligned}$$

□

Section 8.2.7 shows that (8.46) holds precisely for symplectic methods.

We now extend $\bar{u}$ to be a function defined for all t. First extend the solution $\bar{X}$ to all time as a continuous piecewise linear function

$$\bar{X}(t) = \frac{t_{n+1} - t}{\Delta t}\,\bar{X}_n + \frac{t - t_n}{\Delta t}\,\bar{X}_{n+1}, \quad \text{for } t_n \le t < t_{n+1}, \tag{8.47}$$

so that

$$\bar{X}'(t) = H^\delta_\lambda(\bar{\lambda}_{n+1}, \bar{X}_n). \tag{8.48}$$

The following lemma shows that two different solutions can not collide for suitable small Δt.

Lemma 8.29. *There is a positive constant c such that if $\Delta t \le c\delta$ two different solutions $(\bar{X}^1, \bar{\lambda}^1)$ and $(\bar{X}^2, \bar{\lambda}^2)$ of (8.29) do not intersect.*

Proof. Assume there exist two optimal paths $(\bar{X}^1, \bar{\lambda}^1)$ and $(\bar{X}^2, \bar{\lambda}^2)$ that intersect at time t, where $\bar{t}_n < t \le \bar{t}_{n+1}$, then

$$\bar{X}^1_n + (t - \bar{t}_n)H^\delta_\lambda(\bar{\lambda}^1_{n+1}, \bar{X}^1_n) = \bar{X}^2_n + (t - \bar{t}_n)H^\delta_\lambda(\bar{\lambda}^2_{n+1}, \bar{X}^2_n)$$

which can be written

$$\bar{X}^1_n - \bar{X}^2_n = (t - t_n)\big(H^\delta_\lambda(\bar{\lambda}^2_{n+1}, \bar{X}^2_n) - H^\delta_\lambda(\bar{\lambda}^1_{n+1}, \bar{X}^1_n)\big). \tag{8.49}$$

To obtain an estimate of the size of the right hand side in (8.49) integrate along the line

$$\bar{X}(s) = \bar{X}_n^1 + s(\bar{X}_n^2 - \bar{X}_n^1),$$

with $\bar{\lambda}_{n+1}^i$ a function of $\bar{X}_n^i$. The difference in the right hand side of (8.49) is

$$H_\lambda^\delta(\bar{\lambda}_{n+1}^2, \bar{X}_n^2) - H_\lambda^\delta(\bar{\lambda}_{n+1}^1, \bar{X}_n^1) = \int_0^1 \frac{\mathrm{d}H_\lambda^\delta}{\mathrm{d}s}\,\mathrm{d}s$$
$$= \int_0^1 \big(H_{\lambda x}^\delta + H_{\lambda\lambda}^\delta \partial_{\bar{X}_n}\bar{\lambda}_{n+1}\big) ds (\bar{X}_n^2 - \bar{X}_n^1).$$

By assumption it holds that $\|H_{\lambda x}^\delta + H_{\lambda\lambda}^\delta \partial_{\bar{X}_n}\bar{\lambda}_{n+1}\|_{\mathcal{C}} = \mathcal{O}\big(C_\lambda(1+K)/\delta\big)$. Hence the norm of the right hand side in (8.49) is $\mathcal{O}(\delta^{-1}\Delta t)\mathcal{O}\left\|\bar{X}_n^1 - \bar{X}_n^2\right\|$. Therefore there is a positive constant c such that if $\Delta t < c\delta$, the equation (8.49) has only the solution $\bar{X}_n^1 = \bar{X}_n^2$. □

Since the optimal paths $\bar{X}$ do not collide, for suitable small Δt, the value function $\bar{u}$ is uniquely defined along the optimal paths, by (8.31) and

$$\bar{u}\big(\bar{X}(t), t\big) = \bar{u}(\bar{X}_{n+1}, t_{n+1}) + (t_{n+1} - t)h^\delta(\bar{X}_n, \bar{\lambda}_{n+1}), \quad t_n < t < t_{n+1} \tag{8.50}$$

and we are ready for the main lemma

Lemma 8.30. *The value function for the Pontryagin method satisfies a Hamilton-Jacobi equation close to (8.4), more precisely there holds*

$$\begin{aligned} \bar{u}_t + H(\bar{u}_x, \cdot) &= \mathcal{O}(\delta + \Delta t + \frac{(\Delta t)^2}{\delta}) \quad \text{in } \mathbb{R}^d \times (0,T), \\ \bar{u} &= g \quad \text{on } \mathbb{R}^d. \end{aligned} \tag{8.51}$$

The error term $\mathcal{O}(\delta + \Delta t + \frac{(\Delta t)^2}{\delta})$ in (8.51) is a Lipschitz continuous function of $\bar{u}_x(x,t)$, x and t satisfying

$$|\mathcal{O}(\delta + \Delta t + \frac{(\Delta t)^2}{\delta})| \le CC_\lambda\Big(\delta + C_x\Delta t + C_xC_\lambda(1+K)\frac{(\Delta t)^2}{\delta}\Big),$$

where C_x and C_λ are the Lipschitz constants of H in the x and λ variable, respectively, and $C \sim 1$ does not depend on the data.

Proof. The proof starts with the observation

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}t}\bar{u}(\bar{X}(t), t) + h^\delta(\bar{\lambda}_{n+1}, \bar{X}_n) \\ &= \bar{u}_t(\bar{X}(t), t) + \bar{u}_x(\bar{X}(t), t)\cdot f^\delta(\bar{\lambda}_{n+1}, \bar{X}_n) + h^\delta(\bar{\lambda}_{n+1}, \bar{X}_n). \end{aligned} \tag{8.52}$$

The idea is now to use that the dual function $\bar{\lambda}$ is the gradient of $\bar{u}$ at the time levels t_n, by Lemma 8.28, (and a good approximation at times in between) and that the modified discrete Pontryagin method shows that the right hand side in (8.52) is consistent with the correct Hamiltonian H.

We will first derive an estimate of $|\bar{u}_x(\bar{X}(t),t) - \bar{\lambda}_{n+1}|$ for $t_n < t < t_{n+1}$. We have that

$$\bar{u}(\bar{X}(t),t) = \bar{u}(\bar{X}_{n+1},\bar{t}_{n+1}) + (\bar{t}_{n+1} - t)h^\delta(\bar{\lambda}_{n+1},\bar{X}_n)$$

Therefore $\bar{u}_x(\bar{X}(t),t)$ can be written as

$$\begin{aligned}\bar{u}_x(\bar{X}(t),t) &= \frac{\partial \bar{X}_n}{\partial \bar{X}_t}\Big(\frac{\partial \bar{X}_{n+1}}{\partial \bar{X}_n}\bar{u}_x(\bar{X}_{n+1},t_{n+1}) + (t_{n+1}-t)\partial_{\bar{X}_n}h^\delta(\bar{\lambda}_{n+1},\bar{X}_n)\Big)\\ &= \frac{\partial \bar{X}_n}{\partial \bar{X}_t}\Big(\frac{\partial \bar{X}_{n+1}}{\partial \bar{X}_n}\bar{\lambda}_{n+1} + (t_{n+1}-t)\partial_{\bar{X}_n}h^\delta(\bar{\lambda}_{n+1},\bar{X}_n)\Big).\end{aligned}$$

Introduce the notation

$$\begin{aligned}A \equiv \partial_{\bar{X}_n}H^\delta_\lambda(\bar{\lambda}_{n+1},\bar{X}_n) &= H^\delta_{\lambda x}(\bar{\lambda}_{n+1},\bar{X}_n) + H^\delta_{\lambda\lambda}(\bar{\lambda}_{n+1},\bar{X}_n)\partial_{\bar{X}_n}\bar{\lambda}_{n+1}\\ &= \mathcal{O}\big(C_\lambda(1+K)/\delta\big).\end{aligned} \tag{8.53}$$

We have

$$\begin{aligned}\frac{\partial \bar{X}_{n+1}}{\partial \bar{X}_n} &= I + \Delta t A = I + (t-t_n)A + (t_{n+1}-t)A\\ \frac{\partial \bar{X}_n}{\partial \bar{X}_t} &= \big(I + (t-t_n)A\big)^{-1}\end{aligned}$$

therefore as in Lemma 8.28

$$\begin{aligned}&\bar{u}_x(\bar{X}(t),t)\\ &= \bar{\lambda}_{n+1} + (t_{n+1}-t)\big(I+(t-t_n)A\big)^{-1}\big(A\bar{\lambda}_{n+1} + \partial_{\bar{X}_n}h^\delta(\bar{\lambda}_{n+1},\bar{X}_n)\big)\\ &= \bar{\lambda}_{n+1} + (t_{n+1}-t)\big(I+(t-t_n)A\big)^{-1}H^\delta_x(\bar{\lambda}_{n+1},\bar{X}_n)\\ &= \bar{\lambda}_{n+1} + \mathcal{O}\big(C_x\Delta t + C_xC_\lambda(K+1)(\Delta t)^2/\delta\big).\end{aligned} \tag{8.54}$$

Introduce the notation $\tilde{\lambda} \equiv \bar{u}_x(\bar{X}(t),t)$ and split the Hamiltonian term in (8.52) into three error parts:

$$\begin{aligned}r(\tilde{\lambda},\bar{X}(t),t) &\equiv \tilde{\lambda}f^\delta(\bar{\lambda}_{n+1},\bar{X}_n) + h^\delta(\bar{\lambda}_{n+1},\bar{X}_n) - H\big(\tilde{\lambda},\bar{X}(t)\big)\\ &= \tilde{\lambda}f^\delta(\bar{\lambda}_{n+1},\bar{X}_n) + h^\delta(\bar{\lambda}_{n+1},\bar{X}_n) - H^\delta(\tilde{\lambda},\bar{X}_n)\\ &\quad + H^\delta(\tilde{\lambda},\bar{X}_n) - H^\delta\big(\tilde{\lambda},\bar{X}(t)\big)\\ &\quad + H^\delta\big(\tilde{\lambda},\bar{X}(t)\big) - H\big(\tilde{\lambda},\bar{X}(t)\big)\\ &\equiv I + II + III.\end{aligned} \tag{8.55}$$

Taylor expansion of H^δ to second order and (8.54) show

$$\begin{aligned}|I| &= |H^\delta(\bar{\lambda}_{n+1},\bar{X}_n) + (\tilde{\lambda}-\bar{\lambda}_{n+1})H^\delta_\lambda(\bar{\lambda}_{n+1},\bar{X}_n) - H^\delta(\tilde{\lambda},\bar{X}_n)|\\ &\le \min\big(2C_\lambda|\tilde{\lambda}-\bar{\lambda}_{n+1}|, |(\tilde{\lambda}-\bar{\lambda}_{n+1})H^\delta_{\lambda\lambda}(\xi,\bar{X}_n)(\tilde{\lambda}-\bar{\lambda}_{n+1})|/2\big)\\ &\le CC_\lambda\big(C_x\Delta t + C_xC_\lambda(K+1)(\Delta t)^2/\delta\big);\end{aligned}$$

the Lipschitz continuity of H^δ implies

$$|II| \leq |H_x^\delta||\bar{X}(t) - \bar{X}_n| \leq |H_x^\delta||H_\lambda^\delta|\Delta t;$$

and the approximation H^δ satisfies

$$|III| \leq CC_\lambda \delta.$$

The combination of these three estimates proves (8.51).

To finish the proof of the lemma we show that the error function r can be extended to a Lipschitz function in $\mathbb{R}^d \times \mathbb{R}^d \times [0,T]$. We note that by (8.43), (8.47) and (8.54) $\tilde{\lambda}$ is a Lipschitz function of X_t and t, and $r(\tilde{\lambda}(X_t,t),X_t,t)$ is Lipschitz in X_t and t. By

$$r(\lambda, X, t) \equiv r(\tilde{\lambda}(X,t), X, t)$$

we obtain a Lipschitz function r in $\mathbb{R}^d \times \mathbb{R}^d \times [0,T]$.

□

The results in these lemmas finishes the proof of Theorem 8.27: the combination of the residual estimates in Lemma 8.30 and the $\mathcal{C}$-stability estimate of viscosity solutions in Lemma 8.31 proves the theorem.

The approximation result can be extended to the case when the set of backward optimal paths $\{(\bar{X}(t),t) : \ t < T\}$, solving (8.29) and (8.47) , may collide into a codimension one surface Γ in space-time $\mathbb{R}^d \times [0,T]$, see [35].

8.2.3.3 Maximum Norm Stability for Hamilton-Jacobi Equations

The seminal construction of viscosity solutions by Crandall and Lions [9] also includes $\mathcal{C}$ stability results formulated in a general setting. We restate a variant adapted to the convergence results in this paper.

Lemma 8.31. *Suppose $H : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a Lipschitz continuous Hamiltonian satisfying for a constant C and for all $x, \hat{x}, \lambda, \hat{\lambda} \in \mathbb{R}^d$*

$$|H(\lambda,x) - H(\lambda,\hat{x})| \leq C_x|x-\hat{x}|(1+|\lambda|),$$
$$|H(\lambda,x) - H(\hat{\lambda},x)| \leq C_\lambda|\lambda - \hat{\lambda}|.$$

Suppose also that $e : \mathbb{R}^d \times [0,T] \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}$ are Lipschitz continuous. Then, the bounded uniformly continuous viscosity solutions u and $\hat{u}$ of the Hamilton-Jacobi equations

$$u_t + H(u_x, \cdot) = 0 \quad \text{in } \mathbb{R}^d \times (0,T), \quad u|_{\mathbb{R}^d \times \{T\}} = g, \tag{8.56}$$
$$\hat{u}_t + H(\hat{u}_x, \cdot) = e \quad \text{in } \mathbb{R}^d \times (0,T), \quad \hat{u}|_{\mathbb{R}^d \times \{T\}} = g, \tag{8.57}$$

satisfy the $\mathcal{C}$-stability estimate

$$\mathcal{O}\left\|u - \hat{u}\right\|_{\mathcal{C}(\mathbb{R}^d \times [0,T])} \leq T\mathcal{O}\left\|e\right\|_{\mathcal{C}(\mathbb{R}^d \times [0,T])}. \tag{8.58}$$

This follows from the maximum norm stability (8.22), but other proofs based on the maximum principle or the comparison principle are also possible, see [35].

8.2.4 How to obtain the Controls

The optimal control for the exact problem (8.4) is determined by the value function through the Pontryagin principle

$$\alpha(x,t) \in \operatorname*{argmin}_{a\in B}\big(u_x(x,t)\cdot f(x,a)+h(x,a)\big).$$

Assume we have solved a discrete approximating optimal control problem and obtained the approximations $\bar{X}$, $\bar{\lambda}$ and $\bar{u}$. Can they be used to determine an approximation of the control α? Even in the case that the optimal control $S(\lambda,x)\equiv \operatorname{argmin}_a\big(\lambda\cdot f(x,a)+h(x,a)\big)$ is a function, it is in general not continuous as function of x and λ but only piecewise Lipschitz continuous. Therefore the approximate control $S(\bar{\lambda}(t),x)$ cannot be accurate in maximum norm. However, weaker measures of the control can converge; for instance the value function is accurately approximated in Theorems 8.27 and **??**. At the points where S is Lipschitz continuous the error in the control is proportional to the error $|\bar{\lambda}(x,t)-u_x(x,t)|$, for fixed x. If we assume that the error $\bar{u}(\cdot,t)-u(\cdot,t)$ is bounded by ϵ in a $\sqrt{\epsilon}$-neighborhood of x and that $\bar{u}_{xx}$ and u_{xx} also are bounded there, we obtain, for difference quotients $\Delta u/\Delta x$ and $|\Delta x|=\sqrt{\epsilon}$, the error estimate

$$\bar{\lambda}-u_x=\bar{\lambda}-\frac{\Delta\bar{u}}{\Delta x}+\frac{\Delta\bar{u}}{\Delta x}-\frac{\Delta u}{\Delta x}+\frac{\Delta u}{\Delta x}-u_x=\mathcal{O}(\Delta x+\epsilon/\Delta x)=\mathcal{O}(\sqrt{\epsilon}).$$

Convergence of the approximate path $(\bar{X},\bar{\lambda})$ typically requires Lipschitz continuous flux (H_λ,H_x), which we do not assume in this work.

8.2.5 Inverse Problems and Tikhonov Regularization

One way to introduce regularization of ill-posed inverse problems is to study a simple example such as $u'=f$: the forward problem to determine u from f in this case becomes a well-posed integral $u(x)=u(0)+\int_0^x f(s)\mathrm{d}s$ and the inverse problem is then to determine f from u by the derivative $f=u'$. Note that a small error in the data can be amplified when differentiated; for instance a small perturbation maximum-norm $\epsilon\sin(\omega x)$ in u leads to the f-perturbation $\epsilon\omega\cos(\omega x)$ which is large (in maximum-norm) if $\omega\epsilon\gg 1$ even if $\epsilon\ll 1$, while a small maximum-norm perturbation of f leads to a small perturbation of u (in maximum norm). This is the reason that, to determine u from f is well posed (in maximum norm), while the inverse problem to determine f from u is ill posed.

The simplest method to regularize the problem $f=u'$ is to replace the derivative with a difference quotient with suitable step size h. If we assume that our measured values u_* of $u\in\mathcal{C}^2$ are polluted with an error η of size ϵ in maximum norm so that $u_*=u+\eta$, we have

$$f=(u_*-\eta)'.$$

To avoid differentiating η we use the difference quotient

$$\begin{aligned} f(x) &= u'(x) \\ &= \frac{u(x+h)-u(x)}{h} + \mathcal{O}(h) \\ &= \frac{u_*(x+h)-u_*(x)}{h} + \mathcal{O}(\epsilon h^{-1} + h). \end{aligned}$$

The error term is minimal if we choose $h^2 \simeq \epsilon$, that is the optimal step size, $h \simeq \sqrt{\epsilon}$, yields the error $\mathcal{O}(\epsilon^{1/2})$ to compute f by the difference quotient. This difference quotient converges to u' as ϵ tends to zero. If we take too small step size (e.g. $h = \epsilon$), the estimation error does not tend to zero as the measurement error tends to zero.

We can write the inverse problem $u' = f$ as the optimal control problem

$$\begin{aligned} \dot{X}^t &= \alpha^t, \\ \min_{\alpha:(0,1)\to[-M,M]} 2^{-1} \int_0^1 |X^t - X_*^t|^2 \,\mathrm{d}t, \end{aligned}$$

where we changed notation to $t := x$, $X = u$, $X_* = u_*$, $\alpha := f$ and put the constraint to seek α in the bounded set $[-M, M]$ for some positive M. The Hamiltonian becomes

$$H(\lambda, x, t) = \min_{\alpha\in[-M,M]} \left(\lambda \cdot \alpha + 2^{-1}|x - X_*^t|^2\right) = -M|\lambda| + 2^{-1}|x - X_*^t|^2$$

which is not differentiable and leads to the system

$$\begin{aligned} \dot{X}^t &= -M\mathrm{sgn}(\lambda) \\ \dot{\lambda}^t &= -(X^t - X_*^t). \end{aligned}$$

A regularization of this is to replace $\operatorname{sgn}\lambda$ by $\tanh\lambda/\delta$ in the flux, which yields the regularized Hamiltonian

$$H^\delta(\lambda, x, t) = -M\delta \log(\cosh\frac{\lambda}{\delta}) + 2^{-1}|x - X_*^t|^2. \tag{8.59}$$

A standard alternative and related regularization is to add a penalty function depending on the control to the Lagrangian

$$\mathcal{L}^\delta(\lambda, x, \alpha) := \int_0^1 \lambda^t(\alpha^t - \dot{X}^t) + 2^{-1}|X^t - X_*^t|^2 + \delta\alpha^2 \,\mathrm{d}t$$

for some $\delta > 0$, which generates the Hamiltonian system

$$\begin{aligned} \dot{X}^t &= -M\mathrm{sgn}^\delta(\lambda) \\ \dot{\lambda}^t &= -(X^t - X_*^t), \end{aligned}$$

where sgn^δ is the piecewise linear approximation to sgn with slope $-1/(2\delta)$, see Figure 8.13. The corresponding Hamiltonian is $\mathcal{C}^1$ and has the following parabolic approximation

of $-M|\lambda|$

$$\begin{cases} -\lambda M + \delta M^2 & \text{if } \lambda > 2\delta M \\ -\frac{\lambda^2}{4\delta} & \text{if } -2\delta M \le \lambda \le 2\delta M \\ \lambda M + \delta M^2 & \text{if } \lambda \le 2\delta M, \end{cases}$$

which in some sense is the simplest regularization giving a differentiable Hamiltonian. Such a regularization obtained by adding a penalty function, depending on the control, to the Lagranian is called a *Tikhonov regularization*. Any smooth modification of the Hamiltonian can be interpreted as adding such a Tikhonov penalty function, see Section 8.2.5. The fundamental property we desire of a regularization is that the Hamiltonian becomes differentiable. It is somewhat difficult to directly see how to choose a penalty yielding differentiable Hamiltonian, therefore we propose instead to directly regularize the Hamiltonian, e.g. by a mollification as in (8.45) (instead of finding appropriate penalty functions):

- choose a suitable set of controls and its range,
- determine the Hamiltonian,
- mollify the Hamiltonian with a parameter $\delta > 0$ as in (8.45).

Another example of a forward problem is to determine the solution u, representing e.g. temperature, in the boundary value problem

$$\begin{aligned} \big(a(x)u'(x)\big)' &= f(x) \quad 0 < x < 1 \\ u(0) &= u'(1) = 0 \end{aligned} \tag{8.60}$$

for a given source function $f : (0,1) \to (c,\infty)$ and a given conductivity $a : (0,1) \to (c,\infty)$ with $c > 0$. This is a well posed problem with the solution

$$u(x) = \int_0^x \frac{F(s) - F(1)}{a(s)}\,\mathrm{d}s,$$

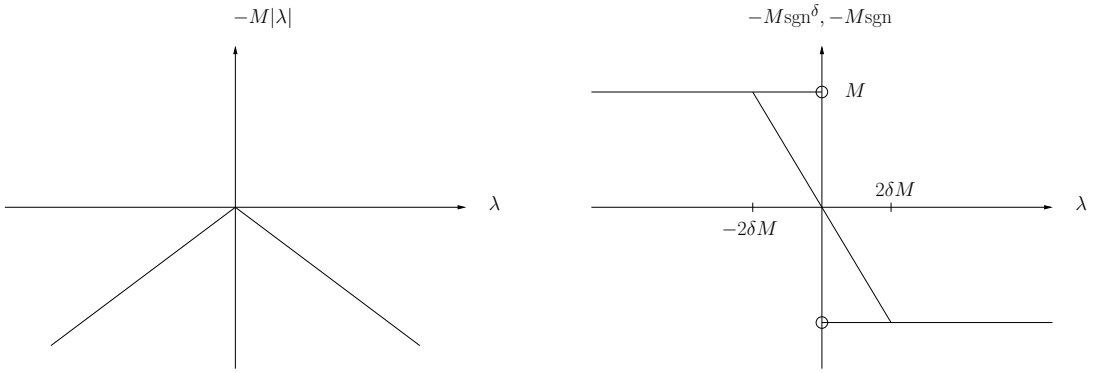


Figure 8.13: Graph of the functions $-M|\lambda|$, $-\mathrm{sgn}^\delta$ and $-\mathrm{sgn}$.

where $F(s) = \int_0^s f(t)\,\mathrm{d}t$ is a primitive function of f. The inverse problem to find the conductivity a from given temperature u and source f leads to

$$a(x) = \frac{F(x) - F(1)}{u'(x)}, \tag{8.61}$$

which depends on the derivative u', as in the previous example, so that it is ill posed (in maximum norm) by the same reason.

Example 8.32 (Numerical regularization)**.** Instead of the exact inversion formula (8.61) we can formulate the optimal control problem

$$\min_{a:[0,1]\to\mathbb{R}} \frac{1}{2}\int_0^1 (u-u^*)^2 + \delta a^2 \mathtt{d}x,$$

where a and x satisfies (8.60), $\delta > 0$, and u^* denotes given data corresponding to a diffusion coefficient a^*. From the Lagrangian

$$\begin{aligned}\mathcal{L}(u,\lambda,a) :=& \frac{1}{2}\int_0^1 (u-u^*)^2 + \delta a^2 + (au')'\lambda - f\lambda \mathtt{d}x = \\ =& \frac{1}{2}\int_0^1 (u-u^*)^2 + \delta a^2 - au'\lambda' - f\lambda \mathtt{d}x,\end{aligned}$$

the Lagrange principle gives that a necessary condition for an optimum is that u, λ and a satisfies Equation (8.60), the dual equation

$$(a(x)\lambda')' = u^* - u, \quad 0 < x < 1, \lambda(0) = \lambda'(1) = 0, \tag{8.62}$$

and

$$u'\lambda' + \delta a = 0, \quad 0 < x < 1. \tag{8.63}$$

In this case the Lagrange principle gives the same result as the Pontryagin principle since the Lagrangian is convex in a, and since it is smooth in a no regularization is needed. For $\delta = 0$, the Pontryagin principle does not give an explicit Hamiltonian unless we impose some bounds on a, while the Lagrange principle still is useful numerically, as we shall see.

The simplest way to solve system (8.60), (8.62) and (8.63) is to use the gradient method: given a starting guess a^i, solve (8.60) to get u, and (8.62) to get λ, and finally update a by taking a step of length θ in the negative gradient direction, i.e.

$$\begin{aligned}a^{i+1} = a^i - \theta\frac{\mathtt{d}\mathcal{L}\big(u(a^i),\lambda(a^i),a^i\big)}{\mathtt{d}a^i} = a^i - \theta\Big(\frac{\partial\mathcal{L}}{\partial u}\frac{\mathtt{d}u}{\mathtt{d}a^i} + \frac{\partial\mathcal{L}}{\partial\lambda}\frac{\mathtt{d}\lambda}{\mathtt{d}a^i} + \frac{\partial\mathcal{L}}{\partial a^i}\Big) = \\ = \Big\{\frac{\partial\mathcal{L}}{\partial u} = \frac{\partial\mathcal{L}}{\partial\lambda} = 0\Big\} = a^i - \theta(u'\lambda' + \delta a^i), \quad 0 < x < 1.\end{aligned}$$

Consider the test problem where the measurement u^* is generated by solving (8.60) with the finite element method for a reference coefficient $a^*(x) := 1 + 0.5\sin(2\pi x)$ and a source term $f = 1$. To the measurements we add some noise, see Figure 8.14.

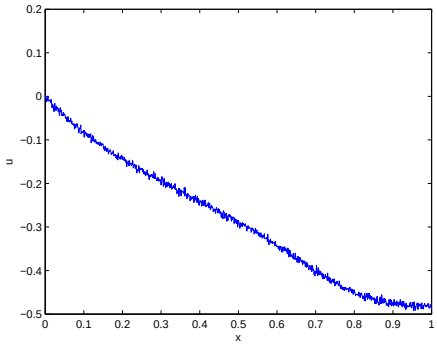


Figure 8.14: Measurements with added noise.

We will now compare different types of regularization: Tikhonov regularization and regularization by discretization or by iteration. In Figure 8.15 the exact inversion (8.61) is shown. A zero misfit error $u - u^*$ here gives an highly oscillating inversion and is thus infeasible for practical use. The only way to use this method is to introduce a numerical regularization from choosing a sufficiently large discretization. In the right part of Figure 8.15 a 100 times coarser mesh is used for the inversion. It is here possible to see something that vaguely resembles the sought coefficient a^*.

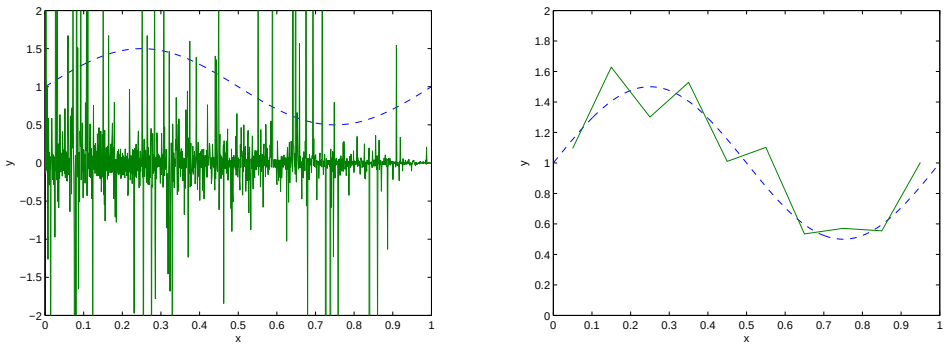


Figure 8.15: Reconstructed coefficient from exact inversion using different meshes.

From the gradient method, for which we choose $\theta = 10$, we can in Figure 8.16 see the result for the case with no noise and $\delta = 0$. Although the absence of noise will theoretically give an exact fit to data, the method will take a long time to converge, and even for a fast method like Newton's method, a small misfit error may still imply a substantial error in the coefficient.

To test the gradient method for the case with measurement noise we start by letting $\delta = 0$. In Figure 8.17 we can see that the gradient method initially finds a smooth function that fits σ^* quite good, but eventually the noise will give a randomly oscillating coefficient as the misfit error decreases. To interrupt the iteration process prematurely is

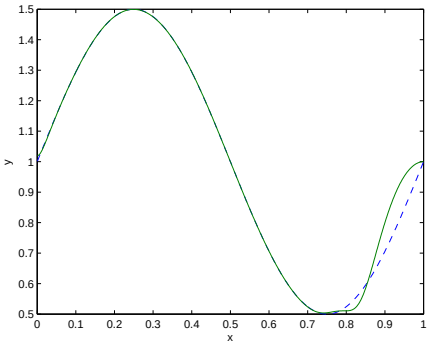


Figure 8.16: Reconstructed coefficient from the gradient method with no noise in measurements and $\delta = 0$.

here a sort of regularization called Landweber iteration [39]. In Figure 8.18 the error in data and coefficients is shown; it is evident that the optimal stopping criterion occurs when the $\|\sigma - \sigma^*\|$ reaches its minimum. Unfortunately, since σ^* is unknown this criterion cannot be fulfilled in practice.

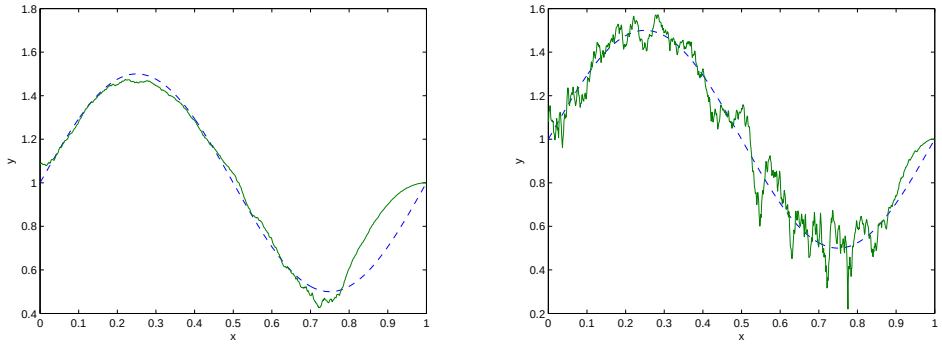


Figure 8.17: Reconstructed coefficient from the gradient method with noisy measurements and $\delta = 0$. Left: 100 iterations. Right: 1000 iterations.

In Figure 8.19 the result for the gradient method with a small regularization $\delta = 5{\cdot}10^{-4}$ is shown. Although the error in the coefficient is higher than for the case with $\delta = 0$, in Figure 8.18, this error is bounded and we can thus continue the iterations until the desired tolerance of the gradient norm is met.

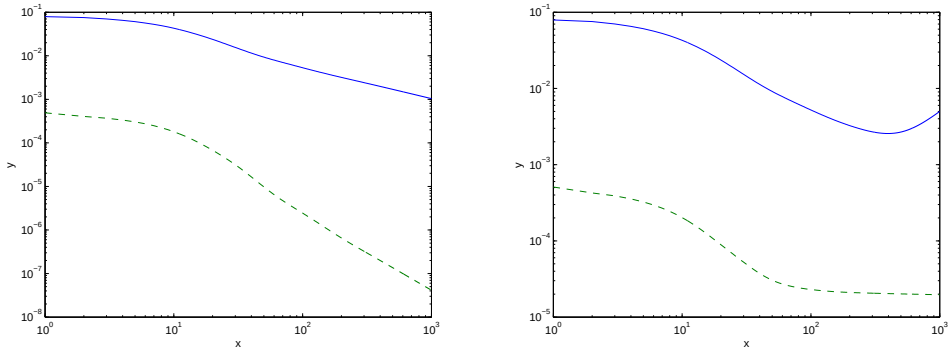


Figure 8.18: Iteration errors from the gradient method. The solid lines depict $\|\sigma - \sigma^*\|^2$ and the dashed lines show $\|u - u^*\|^2$. Left: No noise in data. Right: Noisy data. Note that after a certain number of iterations, $\|\sigma - \sigma^*\|^2$ will get larger as $\|u - u^*\|^2$ gets smaller.

Exercise 8.33. Consider the the following inverse problems:

(i) Estimate a given the solution u to

$$\begin{aligned} \big(a(x)u'(x)\big)' &= 1 \quad 0 < x < 1 \\ u(0) &= u(1) = 0. \end{aligned}$$

(ii) Estimate a given the boundary solution $u(1)$ to

$$\begin{aligned} \big(a(x)u'(x)\big)' &= 0 \quad 0 < x < 1 \\ u(0) &= 0, \\ u'(1) &= 1. \end{aligned}$$

What can we say about the estimation of a for each problem?

Example 8.34. Condition number, matrices, tomography

8.2.6 Smoothed Hamiltonian as a Tikhonov Regularization

The $\mathcal{C}^2$ regularization of the Hamiltonian can also be viewed as a special Tikhonov regularization, using the Legendre transformation: a preliminary idea is to find the Tikhonov penalty function $T(x, \alpha) : \mathbb{R}^d \times A \to \mathbb{R}$ such that

$$\min_{\alpha \in A} \big(\lambda \cdot f(x, \alpha) + T(x, \alpha)\big) = H^\delta(\lambda, x).$$

In general this can only hold if the set A is dense enough, e.g. if A would consist of only two elements the function H^δ would not be smooth. Therefore we replace A seeking the

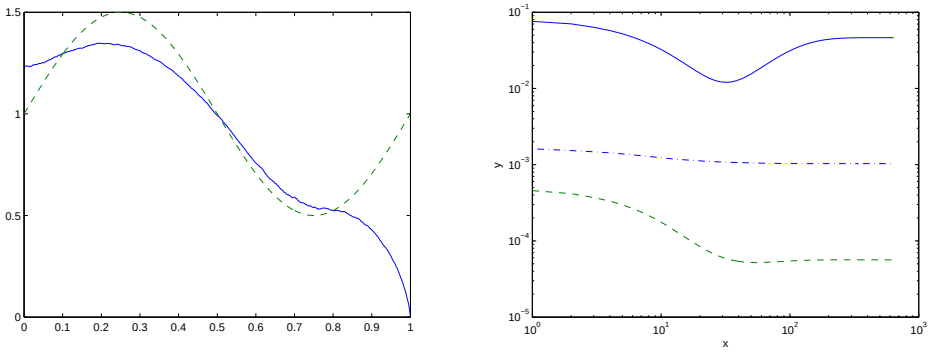


Figure 8.19: Left:Reconstructed coefficient from the gradient method with noisy measurements and $\delta = 5 \cdot 10^{-4}$. Right: Errors as in Figure 8.18 but also including the value function $\|u - u^*\|^2 + \delta\|\sigma - \sigma^*\|^2$ (dash-dotted line).

minimum in the convex closure

$$\widehat{f}(x, A) := \{sf_1 + (1-s)f_2 \mid s \in [0,1], \text{ and } f_1, f_2 \in f(x, A)\}$$

and we instead want to find $T_x(f) : \mathbb{R}^d \times \widehat{f}(x, A) \to \mathbb{R}$ such that

$$\min_{\phi \in \widehat{f}(x,A)} \big(\lambda \cdot \phi + T_x(\phi)\big) = H^\delta(\lambda, x) \quad \text{for all } \lambda \in \mathbb{R}^d. \tag{8.64}$$

To find the Tikhonov penalty, the first step is to observe that by Theorem **??** there is for each λ, where $\partial_\lambda H(\cdot, x)$ is defined, an α such that $\partial_\lambda H(\lambda, x) = f(x, \alpha)$; therefore the regularization $H^\delta(\lambda, x) = \int_{\mathbb{R}^d} H(\lambda - y)\eta(y)dy$, as in (**??**), satisfies $\overline{\partial_\lambda H^\delta(\mathbb{R}^d, x)} \subset \widehat{f}(x, A)$, since H is Lipschitz continuous and hence differentiable almost everywhere.

Define the Legendre transformation

$$\tilde{T}_x(\phi) := \sup_{\lambda \in \mathbb{R}^d} \Big(-\lambda \cdot \phi + H^\delta(\lambda, x)\Big) \quad \text{for all } \phi \in \mathbb{R}^d. \tag{8.65}$$

Figure 8.20 illustrates the value of the Legendre transform

$$T(\phi) = \sup_{\lambda \in \mathbb{R}} \Big(-\lambda \cdot \phi + H(\lambda)\Big)$$

of a concave differentiable function $H : \mathbb{R} \to \mathbb{R}$, i.e. find the tangent to the curve

$$\big\{\big(\lambda, H(\lambda)\big) \mid \lambda \in \mathbb{R}\big\}$$

with the slope ϕ, then its intersection with the y-axis is $T(\phi)$; in multi dimension, $d \geq 1$, find the tangent plane of the graph of H with normal $(\phi, -1)$, then the point $(0, T(\phi))$ is in the plane. If the range of $\partial_\lambda H(\cdot, x)$ is only a subset S of $\mathbb{R}^d$, we see that $T(\phi) = +\infty$ for $\phi \in \mathbb{R}^d - S$.

Theorem 8.35. *By defining $T_x(\phi) := \tilde{T}_x(\phi)$, the relation (8.64) holds.*

Proof. Fix a point $x \in \mathbb{R}^d$. The definition (8.65) of the Legendre transform implies that for any ϕ and all $\lambda \in \mathbb{R}^d$ we have

$$\lambda \cdot \phi + \tilde{T}_x(\phi) \geq H^\delta(\lambda, x). \tag{8.66}$$

It remains to show that for any λ we can have equality here by choosing ϕ precisely.

Since the Hamiltonian $H^\delta(\cdot, x)$ is concave and differentiable, with $\partial_\lambda H^\delta(\cdot, x) \in \widehat{f}(x, A)$, the maximum in the Legendre transform is, for ϕ in the interior of $\widehat{f}(x, A)$, attained at a point λ^* (depending on ϕ) satisfying

$$\tilde{T}_x(\phi) = \sup_{\lambda \in \mathbb{R}^d} \big(-\lambda \cdot \phi + H^\delta(\lambda, x)\big) = -\lambda^* \cdot \phi + H^\delta(\lambda^*, x)$$

and $\phi = \partial_\lambda H^\delta(\lambda^*, x)$, so that the choise $\phi = \partial_\lambda H^\delta(\lambda, x)$ gives equality in (8.66). The fact that $\tilde{T}_x$ is lower semicontinuous shows that

$$\inf_{\phi \in \text{ interior}\widehat{f}(x,A)} \big(\lambda \cdot \phi + \tilde{T}_x(\phi)\big) = \min_{\phi \in \widehat{f}(x,A)} \big(\lambda \cdot \phi + \tilde{T}_x(\phi)\big).$$

□

Exercise 8.36. Show that Tikhonov penalty for the regularized Hamiltonian (8.59) in the $u' = f$ problem is

$$\frac{\delta M^2}{2}\Big((1 + \frac{\alpha}{M})\log((1 + \frac{\alpha}{M}) + (1 - \frac{\alpha}{M})\log(1 - \frac{\alpha}{M})\Big) + \frac{1}{2}|x - X_*^t|^2.$$

8.2.7 General Approximations

The essential property of the symplectic Euler method we have used is that $\bar{u}_x(\bar{X}_n, t_n) = \bar{\lambda}_n$. This relation holds precisely for symplectic approximations (cf. Remark 8.38):

Theorem 8.37. *Consider a general one step method*

$$\begin{aligned} \bar{X}_{n+1} &= A(\bar{\lambda}_{n+1}, \bar{X}_n) \\ \bar{\lambda}_n &= C(\bar{\lambda}_{n+1}, \bar{X}_n) \end{aligned} \tag{8.67}$$

with

$$\bar{u}(\bar{X}_n, t_n) = g(\bar{X}_N) + \sum_{m=n}^{N-1} B(\bar{\lambda}_{n+1}, \bar{X}_n)\Delta t.$$

Then $\bar{u}_x(\bar{X}_n, t_n) = \bar{\lambda}_n$, for all n, implies that the mapping $\phi : (\bar{X}_n, \bar{\lambda}_n) \mapsto (\bar{X}_{n+1}, \bar{\lambda}_{n+1})$ is symplectic. If ϕ is symplectic it is possible to choose the function B so that $\bar{u}_x(\bar{X}_n, t_n) = \bar{\lambda}_n$, for all n.

Proof. As in Lemma 8.28 we have

$$\bar{u}_x(\bar{X}_n, t_n) = \frac{dA\big(\bar{X}_n, \bar{\lambda}_{n+1}(\bar{X}_n)\big)}{d\bar{X}_n}\bar{u}_x(\bar{X}_{n+1}, t_{n+1}) + \frac{dB\big(\bar{X}_n, \bar{\lambda}_{n+1}(\bar{X}_n)\big)}{d\bar{X}_n}.$$

Therefore the relation

$$\bar{u}_x(\bar{X}_n, t_n) = \bar{\lambda}_n$$

holds if and only if $\lambda A_\lambda + B_\lambda = 0$ and $\lambda A_x + B_x = C$. Let $S \equiv \lambda A + B$. Then $\lambda A_\lambda + B_\lambda = 0$ is equivalent to $S_\lambda = A$, but $S_\lambda = A$ implies $B = S - \lambda S_\lambda$ so that $\lambda A_x + B_x = S_x$. Therefore $\lambda A_\lambda + B_\lambda = 0$ and $\lambda A_x + B_x = C$ is equivalent to $A = S_\lambda$ and $C = S_x$.

Let $S \equiv \bar{\lambda}_{n+1} \cdot \bar{X}_n + \Delta t \tilde{H}(\bar{\lambda}_{n+1}, \bar{X}_n)$. Then (8.67), with $A = S_\lambda$ and $C = S_x$, becomes

$$\begin{aligned} \bar{X}_{n+1} &= \bar{X}_n + \Delta t \tilde{H}_\lambda(\bar{X}_n, \bar{\lambda}_{n+1}) \\ \bar{\lambda}_n &= \bar{\lambda}_{n+1} + \Delta t \tilde{H}_x(\bar{X}_n, \bar{\lambda}_{n+1}), \end{aligned} \tag{8.68}$$

which by Remark 8.38 is equivalent to symplecticity of the mapping $(\bar{X}_n, \bar{\lambda}_n) \mapsto (\bar{X}_{n+1}, \bar{\lambda}_{n+1})$. □

Remark 8.38. A one step method (8.67), interpreted as

$$(\bar{X}_n, \bar{\lambda}_n) \mapsto (\bar{X}_{n+1}, \bar{\lambda}_{n+1}),$$

is called symplectic if there exists a function $\tilde{H}(\bar{\lambda}_{n+1}, \bar{X}_n)$ such that (8.68) holds, see Theorem 5.1, Lemma 5.2 and (5.5) in Chapter VI of [20], where a thorough study on symplectic methods can be found.

To generalize the error estimate of Theorems 8.27 and **??** to general symplectic one step approximations (8.68), e.g. the second order symplectic Runge-Kutta method

$$\tilde{H} = \frac{1}{2}\Big(H(\bar{\lambda}_{n+1}, \bar{X}_n) + H\big(\bar{\lambda}_{n+1} + \Delta t H_x(\bar{\lambda}_{n+1}, \bar{X}_n), \bar{X}_n + \Delta t H_\lambda(\bar{\lambda}_{n+1}, \bar{X}_n)\big)\Big)$$

requires first an extension of $\bar{X}_n$ and $\bar{u}$ to all time, by approximations $(\bar{f}, \bar{h})$ of (f^δ, h^δ) with

$$\frac{\mathrm{d}\bar{X}}{\mathrm{d}t} = \bar{f} \quad \text{and} \quad \frac{\mathrm{d}\bar{u}}{\mathrm{d}t} = -\bar{h},$$

and then an estimate of the residual error r as in (8.55). In practice we need more regularity of H^δ to take advantage of higher order methods. Since we only have Lipschitz bounds of H the estimate of r is not smaller than the error $h^\delta - \bar{h}$, which is $\mathcal{O}(\|H^\delta\|_{\mathcal{C}^p})(\Delta t)^p = \mathcal{O}((\Delta t)^p/\delta^{p-1})$ for a pth order accurate method. Consequently the residual error is not smaller than $\mathcal{O}(\delta + (\Delta t)^p/\delta^{p-1}) = \mathcal{O}(\Delta t)$ for $\delta \simeq \Delta t$, so that our error estimate does not improve for higher order schemes, without additional assumptions. On the other hand by extending $\bar{X}$ as a piecewise linear function, as before, the only change of the analysis in Sections 8.2.3.2 and **??** to other symplectic methods (8.68) is to replace $H^\delta(\bar{\lambda}_{n+1}, \bar{X}_n)$ by $\tilde{H}(\bar{\lambda}_{n+1}, \bar{X}_n)$ and since

$$\|H^\delta - \tilde{H}\|_{\mathcal{C}} + \delta\|H^\delta - \tilde{H}\|_{\mathcal{C}^1} + \delta^2\|H^\delta - \tilde{H}\|_{\mathcal{C}^2} = \mathcal{O}(\Delta t)$$

the estimate (8.51) holds for all symplectic methods which are at least first order accurate.

Similarly, by considering $(\bar{X}_{n+1}, \bar{\lambda}_n)$, instead of $(\bar{X}_n, \bar{\lambda}_{n+1})$, as independent variables the scheme

$$\begin{aligned} \bar{X}_n &= A(\bar{X}_{n+1}, \bar{\lambda}_n) \\ \bar{\lambda}_{n+1} &= C(\bar{X}_{n+1}, \bar{\lambda}_n), \end{aligned}$$

is symplectic if and only if

$$\begin{aligned} \bar{X}_n &= \bar{X}_{n+1} - \Delta t \hat{H}_\lambda(\bar{X}_{n+1}, \bar{\lambda}_n) \\ \bar{\lambda}_{n+1} &= \bar{\lambda}_n - \Delta t \hat{H}_x(\bar{X}_{n+1}, \bar{\lambda}_n), \end{aligned} \tag{8.69}$$

and the error analysis of the methods (8.68) applies with

$$\tilde{H}(\bar{X}_n, \bar{\lambda}_{n+1}) = (\bar{X}_{n+1} - \bar{X}_n) \cdot (\bar{\lambda}_{n+1} - \bar{\lambda}_n) + \hat{H}(\bar{X}_{n+1}, \bar{\lambda}_n).$$

An example of a method (8.69) is the Euler method $\hat{H} = H$, which is backward Euler for $\bar{X}$ forwards in time and backward Euler for $\bar{\lambda}$ backwards in time, in contrast to (8.29) which is forward Euler for $\bar{X}$ forwards in time and forward Euler for $\bar{\lambda}$ backwards in time.

8.3 Optimal Control of Stochastic Differential Equations

In this section we study optimal control of the solution $X(t)$ to the stochastic differential equation

$$\begin{cases} dX_i &= a_i(X(s), \alpha(s, X(s)))dt + b_{ij}(X(s), \alpha(s, X(s)))dW_j, \quad t < s < T \\ X(t) &= x \end{cases} \tag{8.70}$$

where T is a fixed terminal time and $x \in \mathbb{R}^n$ is a given initial point. Assume that $a_i, b_{ij} : \mathbb{R}^n \times A \to \mathbb{R}$ are smooth bounded functions, where A is a given compact subset of $\mathbb{R}^m$. The function $\alpha : [0, T] \times \mathbb{R}^n \to A$ is a *control* and let $\mathcal{A}$ be the set of admissible Markov control functions $t \to \alpha(t, X(t))$. The Markov control functions use the current value $X(s)$ to affect the dynamics of X by adjusting the drift and the diffusion coefficients. Let us for these admissible controls $\alpha \in \mathcal{A}$ define the *cost*

$$C_{t,x}(\alpha) = E[\int_t^T h(X(s), \alpha(s))ds + g(X(T))]$$

where X solves the stochastic differential equation (8.70) with control α and

$$h : \mathbb{R}^n \times A \to \mathbb{R}, \quad g : \mathbb{R}^n \to \mathbb{R}$$

are given smooth bounded functions. We call h the *running cost* and g the *terminal cost*. Our goal is to find an optimal control α^* which minimizes the expected cost, $C_{t,x}(\alpha)$.

Let us define the value function

$$u(t, x) \equiv \inf_{\alpha \in \mathcal{A}} C_{t,x}(\alpha). \tag{8.71}$$

The plan is to show that u solves a certain Hamilton-Jacobi equation and that the optimal control can be reconstructed from u. We first assume for simplicity that the optimal control is attained, i.e

$$u(t,x) = \min_{\alpha \in \mathcal{A}} C_{t,x}(\alpha) = C_{t,x}(\alpha^*).$$

The generalization of the proofs without this assumption is discussed in Exercise 8.45.

8.3.1 An Optimal Portfolio

Example 8.39. Assume that the value of a portfolio, $X(t)$, consists of risky stocks, $S(t) = \alpha(t)X(t)$, and risk less bonds, $B(t) = (1-\alpha(t))X(t)$, where $\alpha(t) \in [0,1]$ and

$$dS = aSdt + cSdW, \tag{8.72}$$
$$dB = bBdt, \tag{8.73}$$

with $0 \leq b < a$. Define for a given function g the cost function

$$C_{t,x}(\alpha) = E[g(X(T))|X(t) = x].$$

Then our goal is to determine the Markov control function $\alpha(t, X(t))$, with $\alpha : [0,T] \times \mathbb{R} \to [0,1]$ that maximizes the cost function. The solution will be based on the function

$$u(t,x) \equiv \max_{\alpha} C_{t,x}(\alpha),$$

and we will show that $u(t,x)$ satisfies the following *Hamilton-Jacobi* equation,

$$u_t + \max_{\alpha \in [0,1]} \left\{ (a\alpha + b(1-\alpha))xu_x + \frac{c^2\alpha^2}{2}x^2u_{xx} \right\} = 0, \tag{8.74}$$
$$u(T,x) = g(x),$$

that is

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

for

$$H(x,p,w) \equiv \max_{v \in [0,1]} (av + b(1-v)xp + \frac{c^2v^2}{2}x^2w).$$

Example 8.40. Assume that $u_{xx} < 0$ in the equation (8.74). Determine the optimal control function α^*.

Solution. By differentiating $f(\alpha) = (a\alpha + b(1-\alpha))xu_x + \frac{c^2\alpha^2}{2}x^2u_{xx}$ in (8.74) with respect to α and using $df/d\alpha = 0$, we obtain

$$\hat{\alpha} = -\frac{(a-b)u_x}{c^2xu_{xx}}.$$

Then the optimal control α^* is given by

$$\alpha^* = \begin{cases} 0, & if\ \hat{\alpha} < 0 \\ \hat{\alpha}, & if\ \hat{\alpha} \in [0,1] \\ 1 & if\ 1 < \hat{\alpha} \end{cases}$$

The optimal value yields in (8.74) the Hamilton-Jacobi equation

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

where

$$H(x, u_x, u_{xx}) = \begin{cases} bxu_x, \text{ if } \hat{\alpha} < 0 \\ bxu_x - \dfrac{(a-b)^2 u_x^2}{2c^2 u_{xx}}, \text{ if } \hat{\alpha} \in [0,1] \\ axu_x + \dfrac{c^2 x^2 u_{xx}}{2} \text{ if } 1 < \hat{\alpha} \end{cases} \tag{8.75}$$

□

Example 8.41. What is the optimal control function $\alpha = \alpha^*$ for $g(x) = x^r, 0 < r < 1$?

Solution. We have $dX = d(\alpha X + (1-\alpha)X) = dS + dB = (aS + bB)dt + cSdW = (a\alpha X + b(1-\alpha)X)dt + c\alpha X dW$, so that the Itô formula yields

$$\begin{aligned} dg(X) &= dX^r = rX^{r-1}dX + \frac{r(r-1)}{2}X^{r-2}(dX)^2 \\ &= rX^r(a\alpha + b(1-\alpha))dt + rX^r \alpha c dW + \frac{1}{2}\alpha^2 c^2 r(r-1)X^r dt. \end{aligned}$$

Taking the expectation value in the above,

$$E[X^r(T)] = X^r(0) + E\left[\int_0^T rX^r \left(a\alpha + b(1-\alpha) + \frac{1}{2}\alpha^2 c^2 (r-1)\right) dt\right].$$

Finally, perturb the above equation with respect to $\epsilon \in \mathbb{R}_+$ provided $\alpha = \alpha^* + \epsilon v$ for some feasible function v, that is $\alpha^* + \epsilon v \in [0,1]$ for sufficiently small ϵ. Then the optimal control, α^*, should satisfy $E[X^r_{\alpha^*+\epsilon v}(T)] - E[X^r_{\alpha^*}(T)] \leq 0\ \forall v$. If we make the assumption $\alpha^* \in (0,1)$, then we obtain

$$E[\int_0^T rX^r v(a - b + \alpha^* c^2 (r-1))dt] = 0,\ \forall v$$

which implies

$$\alpha^* = \frac{a-b}{c^2(1-r)}.$$

□

Exercise 8.42. What is the optimal control in (8.74) for $g(x) = \log x$?

8.3.2 Dynamic Programming and Hamilton-Jacobi Equations

Lemma 8.43. *Assume that the assumptions in section 8.3.1 hold. Then, the function u satisfies, for all $\delta > 0$, the dynamic programming relation*

$$u(t,x) = \min_{\alpha:[t,t+\delta]\to A} E[\int_t^{t+\delta} h(X(s),\alpha(s,X(s)))ds + u(t+\delta, X(t+\delta))]. \tag{8.76}$$

Proof. The proof has two steps: to use the optimal control to verify

$$u(t,x) \geq \min_{\alpha\in\mathcal{A}} E[\int_t^{t+\delta} h(X(s),\alpha(s))ds + u(t+\delta, X(t+\delta))],$$

and then to show that an arbitrary control yields

$$u(t,x) \leq \min_{\alpha\in\mathcal{A}} E[\int_t^{t+\delta} h(X(s),\alpha(s))ds + u(t+\delta, X(t+\delta))],$$

which together imply Lemma 8.43.

Step 1: Choose the optimal control α^*, from t to T, to obtain

$$\begin{aligned}
u(t,x) &= \min_{\alpha\in\mathcal{A}} E\left[\int_t^T h(X(s),\alpha(s,X(s)))ds + g(X(T))\right] \\
&= E[\int_t^{t+\delta} h(X(s),\alpha^*(s))ds] + E[\int_{t+\delta}^T h(X(s),\alpha^*(s))ds + g(X(T))] \\
&= E[\int_t^{t+\delta} h(X(s),\alpha^*(s))ds] \\
&\quad + E\left[E[\int_{t+\delta}^T h(X(s),\alpha^*(s))ds + g(X(T))|\ X(t+\delta)]\right] \\
&\geq E[\int_t^{t+\delta} h(X(s),\alpha^*(s))ds] + E[u(X(t+\delta),t+\delta)] \\
&\geq \min_{\alpha\in\mathcal{A}} E\left[\int_t^{t+\delta} h(X(s),\alpha(s,X(s))ds + u(X(t+\delta),t+\delta)\right].
\end{aligned}$$

Step 2: Choose the control α^+ to be arbitrary from t to $t+\delta$ and then, given the value $X(t+\delta)$, choose the optimal α^* from $t+\delta$ to T. Denote this control by $\alpha' = (\alpha^+, \alpha^*)$.

Definition (8.71) shows

$$\begin{aligned} u(t,x) &\leq C_{t,x}(\alpha') \\ &= E[\int_t^T h(X(s),\alpha'(s))ds + g(X(T))] \\ &= E[\int_t^{t+\delta} h(X(s),\alpha^+(s))ds] + E[\int_{t+\delta}^T h(X(s),\alpha^*(s))ds + g(X(T))] \\ &= E[\int_t^{t+\delta} h(X(s),\alpha^+(s))ds] \\ &\quad + E\left[E[\int_{t+\delta}^T h(X(s),\alpha^*(s))ds + g(X(T))|\ X(t+\delta)]\right] \\ &= E[\int_t^{t+\delta} h(X(s),\alpha^+(s))ds] + E[u(X(t+\delta),t+\delta)]. \end{aligned}$$

Taking the minimum over all controls α^+ yields

$$u(t,x) \leq \min_{\alpha^+\in\mathcal{A}} E\left[\int_t^{t+\delta} h(X(s),\alpha^+(s))ds + u(X(t+\delta),t+\delta)\right].$$

□

Theorem 8.44. *Assume that X solves (8.70) with a Markov control function α and that the function u defined by (8.71) is bounded and smooth. Then u satisfies the Hamilton-Jacobi equation*

$$\begin{aligned} &u_t + H(t,x,Du,D^2u) = 0, \\ &u(T,x) = g(x), \end{aligned}$$

with the Hamiltonian function

$$H(t,x,Du,D^2u) \equiv \min_{\alpha\in A}\left[a_i(x,\alpha)\partial_{x_i}u(t,x) + \frac{b_{ik}(x,\alpha)b_{jk}(x,\alpha)}{2}\partial_{x_ix_j}u(t,x) + h(x,\alpha)\right]$$

Proof. The proof has two steps: to show that the optimal control $\alpha = \alpha^*$ yields

$$u_t + a_i^*\partial_{x_i}u + \frac{b_{ik}^*b_{jk}^*}{2}\partial_{x_ix_j}u + h^* = 0, \tag{8.77}$$

where $a^*(x) = a(x,\alpha^*(t,x)), b^*(x) = b(x,\alpha^*(t,x))$ and $h^*(t,x) = h(t,x,\alpha^*(t,x))$, and then that an arbitrary control α^+ implies

$$u_t + a_i^+\partial_{x_i}u + \frac{b_{ik}^+b_{jk}^+}{2}\partial_{x_ix_j}u + h^+ \geq 0, \tag{8.78}$$

where $a^+(x) = a(x,\alpha^+(t,x)), b^+(x) = b(x,\alpha^+(t,x))$ and $h^+(t,x) = h(t,x,\alpha^+(t,x))$. The two equations (8.77) and (8.78) together imply Theorem 8.44.

Step 1 : Choose $\alpha = \alpha^*$ to be the optimal control in (8.70). Then by the dynamic programming principle of Lemma 8.71

$$u(X(t),t) = E[\int_t^{t+\delta} h(X(s),\alpha^*(s,X(s)))ds + u(X(t+\delta),t+\delta)],$$

so that Itô 's formula implies

$$\begin{aligned} -h(t,x,\alpha^*(t,x))dt &= E[du(X(t),t)|\ X(t)=x] \\ &= (u_t + a_i^*\partial_{x_i}u + \frac{b_{ik}^*b_{jk}^*}{2}\partial_{x_ix_j}u)(t,x)dt. \end{aligned} \tag{8.79}$$

Definition (8.71) shows

$$u(T,x) = g(x),$$

which together with (8.79) prove (8.77).

Step 2 : Choose the control function in (8.70) to be arbitrary from time t to $t+\delta$ and denote this choice by $\alpha = \alpha^+$. The function u then satisfies by Lemma 8.71

$$u(t,x) \leq E[\int_t^{t+\delta} h(X(s),\alpha^+(s))ds] + E[u(X(t+\delta),t+\delta)].$$

Hence $E[du] \geq -h(x,\alpha^+)dt$. We know that for any given α^+, by Itô 's formula,

$$E[du(t,X(t))] = E\left[u_t + a_i^+\partial_{x_i}u + \frac{b_{ik}^+b_{jk}^+}{2}\partial_{x_ix_j}u\right]dt.$$

Therefore, for any control α^+,

$$u_t + a_i^+\partial_{x_i}u + \frac{b_{ik}^+b_{jk}^+}{2}\partial_{x_ix_j}u + h(x,\alpha^+) \geq 0,$$

which proves (8.78) □

Exercise 8.45. Use a minimizing sequence α_i of controls, satisfying

$$u(t,x) = \lim_{i\to\infty} C_{t,x}(\alpha_i),$$

to prove Lemma 8.71 and Theorem 8.44 without the assumption that the minimum control is attained.

Exercise 8.46. Let $\mathcal{A}^+$ be the set of all adapted controls $\{\alpha : [0,T]\times\mathcal{C}[0,T] \to A\}$ where $\alpha(s,X)$ may depend on $\{X(\tau) : \tau \leq s\}$. Show that the minimum over all adapted controls in $\mathcal{A}^+$ is in fact the same as the minimum over all Markov controls, that is

$$\inf_{\alpha\in\mathcal{A}^+} C_{t,x}(\alpha) = \inf_{\alpha\in\mathcal{A}} C_{t,x}(\alpha),$$

e.g. by proving the dynamic programming relation (8.76) for adapted controls and motivate why this is sufficient.

8.3.3 Relation of Hamilton-Jacobi Equations and Conservation Laws

In this section we will analyze qualitative behavior of Hamilton-Jacobi equations, in particular we will study the limit corresponding to vanishing noise in control of stochastic differential equations. The study uses the relation between the Hamilton-Jacobi equation for $V : [0,T] \times \mathbb{R} \to \mathbb{R}$

$$V_t + H(V_x) = 0, \quad V(0,x) = V_0(x), \qquad (H-J)$$

and the conservation law for $U : [0,T] \times \mathbb{R} \to \mathbb{R}$

$$U_t + H(U)_x = 0, \quad U(0,x) = U_0(x). \qquad (C-L)$$

Observe that the substitution $V(t,x) = \int_{-\infty}^{x} U(t,y)dy$, so that $U = V_x$, and integration in x from $-\infty$ to x in (C-L) shows

$$V_t + H(V_x) = H(U(t,-\infty)). \tag{8.80}$$

Combined with the assumptions $U(t,x) \to 0$ as $|x| \to \infty$ and $H(0) = 0$ we conclude that V solves (H-J), if U solves (C-L).

The next step is to understand the nature of the solutions of (C-L). Consider the special Burger's conservation law

$$0 = U_t + U\ U_x = U_t + (\frac{U^2}{2})_x, \quad U(0,x) = U_0(x). \tag{8.81}$$

Let us define a *characteristic path* $X : [0,T] \times \mathbb{R} \to \mathbb{R}$ by

$$\frac{dX}{dt}(t) = U(t, X(t)), \ X(0) = x_0. \tag{8.82}$$

Thus, if $\psi(t) \equiv U(t, X(t))$ then $\frac{d\psi}{dt}(t) = 0$ by virtue of (8.81). This means that the value of U is constant along a characteristic path. If the characteristics do not collide into each other we may expect to find a solution using the initial data $U_0(x)$ and the set of characteristics. Unfortunately, this is not what happens in general, and collisions between characteristics do exist and give birth to discontinuities known as shocks. For example, this is the case when $U_0(x) = -\arctan(x)$ and $t \geq 1$.

Exercise 8.47. Show that $w(t) = U_x(X(t),t)$ satisfies $w(t) = w(0)/(1+w(0)t)$, $t < 1$, for Burger's equation (8.81) with initial data $U(x,0) = -\arctan(x)$. Hence, $w(1) = \infty$, for $X(0) = 0$.

Since the method of characteristics does not work globally we have to find an alternative way to explain what happens with the solution $U(t,x)$ near a shock. It is not enough with the concept of strong or classical solution, since the solution $U(t,x)$ is not differentiable in general. For this purpose, we define the notion of weak solution. Let V be the set of test functions $\{\varphi : (0,+\infty) \times \mathbb{R} \to \mathbb{R}\}$ which are differentiable and take the

value zero outside some compact set. Then an integrable function U is a weak solution of (8.81) if it satisfies

$$\int_0^{+\infty}\int_{-\infty}^{+\infty}\left(U(t,x)\varphi_t(t,x)+\frac{U^2(t,x)}{2}\varphi_x(t,x)\right)dx\ dt=0,\ \forall\varphi\in V \tag{8.83}$$

and

$$\int_{-\infty}^{+\infty}|U(t,x)-U_0(x)|dx\to 0,\quad as\ t\to 0 \tag{8.84}$$

Example 8.48. The shock wave

$$U(t,x)=\begin{cases}1 & x<\frac{t}{2},\\ 0 & \text{otherwise.}\end{cases}$$

is a weak solution satisfying (8.83) and (8.84). Observe that for $s\equiv 1/2$

$$\partial_t\int_a^b U\ dx=\frac{U^2(t,a)-U^2(t,b)}{2}=-\left[\frac{U^2}{2}\right],$$

and

$$\partial_t\int_a^b U\ dx=\partial_t\big((s\ t-a)U_-]+(b-s\ t)U_+\big)=-s(U_+-U_-),$$

where

$$[w(x_0)]\equiv w_+(x_0)-w_-(x_0)\equiv\lim_{y\to 0+}w(x_0+y)-w(x_0-y)$$

is the jump at the point x_0. Consequently, the speed s of a shock can be determined by the so called *Rankine Hugoniot* condition

$$s[U]=\left[\frac{U^2}{2}\right]. \tag{8.85}$$

Exercise 8.49. Verify that the shock wave solution

$$U_I(t,x)=\begin{cases}0 & x>-\frac{t}{2},\\ -1 & \text{otherwise}\end{cases}$$

and the rarefaction wave solution

$$U_{II}(t,x)=\begin{cases}0 & x\geq 0,\\ \frac{x}{t} & -t<x<0,\\ -1 & \text{otherwise}\end{cases}$$

are both weak solutions of $U_t+U\ U_x=0$ with the same initial condition.

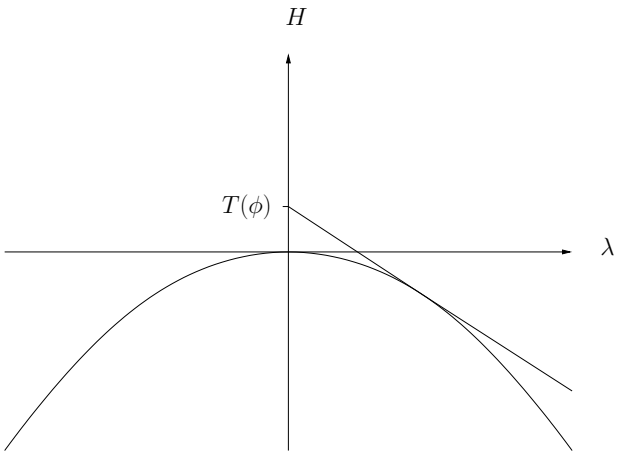


Figure 8.20: Illustration of the Legendre transform. If H decreases sufficiently fast as $|\lambda| \to \infty$, then $\partial_\lambda H$ can attain all values in $\mathbb{R}$ and the range of T is $[0, \infty)$, since $T(0) = 0$ here. If, on the other hand, the slope of H is in an interval I, then $T(I) = [0, T_+)$ for some upper bound T_+, and $T(\mathbb{R} - I) = \{+\infty\}$.

Figure 8.21: Left: Initial condition. Right: Colliding characteristics and a shock.

Figure 8.22: Shock velocity and Rankine Hugoniot condition

Figure 8.23: $U_I(t, x)$

Figure 8.24: $U_{II}(t, x)$

The last exercise shows that we pay a price to work with weak solutions: the lack of uniqueness. Therefore, we need some additional physical information to determine a unique weak solution. This leads us to the concept of *viscosity limit* or *viscosity solution*: briefly, it says that the weak solution U we seek is the limit $U = \lim_{\epsilon \to 0+} U^\epsilon$ of the solution of the regularized equation

$$U_t^\epsilon + U^\epsilon \, U_x^\epsilon = \epsilon U_{xx}^\epsilon, \quad \epsilon > 0. \tag{8.86}$$

This regularized equation has continuous and smooth solutions for $\epsilon > 0$. With reference to the previous example, the weak solution U_{II} satisfies $U_{II} = \lim_{\epsilon \to 0^+} U^\epsilon$, but $U_I \neq \lim_{\epsilon \to 0+} U^\epsilon$. Since a solution of the conservation law can be seen as the derivative of the solution of a Hamilton-Jacobi equation, the same technique of viscosity solutions can be applied to

$$V_t^\epsilon + \frac{(V_x^\epsilon)^2}{2} = \epsilon V_{xx}^\epsilon, \quad \epsilon > 0. \tag{8.87}$$

The functions $V_I(x,t) = -\int_x^\infty U_I(y,t)dy$, and $V_{II}(x,t) = -\int_x^\infty U_{II}(y,t)dy$ have the same initial data and they are both candidates of solutions to the Hamilton-Jacobi equation

$$V_t + \frac{(V_x)^2}{2} = 0.$$

The shock waves for conservation laws corresponds to solutions with discontinuities in the derivative for Hamilton-Jacobi solutions. Only the function V_{II} satisfies

$$V_{II} = \lim_{\epsilon \to 0+} V^\epsilon, \tag{8.88}$$

but $V_I \neq \lim_{\epsilon \to 0+} V^\epsilon$. It can be shown that the condition (8.88) implies uniqueness for Hamilton-Jacobi equations. Note that (8.88) corresponds to the limit of vanishing noise in control of stochastic differential equations.

8.3.4 Numerical Approximations of Conservation Laws and Hamilton-Jacobi Equations

We have seen that the viscous problem

$$\begin{aligned} &\partial_t u^\varepsilon + \partial_x H(u^\varepsilon) = \varepsilon u_{xx}^\varepsilon \quad \text{for} \quad (x,t) \in \mathbb{R} \times (0, +\infty), \\ &u^\varepsilon(x,0) = u_0(x) \qquad \text{for} \quad x \in \mathbb{R}, \end{aligned} \tag{8.89}$$

can be used to construct unique solutions to the conservation law

$$\begin{aligned} &\partial_t u + \partial_x H(u) = 0 \quad \text{for} \quad (x,t) \in \mathbb{R} \times (0, +\infty), \\ &u(x,0) = u_0(x) \qquad \text{for} \quad x \in \mathbb{R}. \end{aligned} \tag{8.90}$$

In this section we will develop numerical approximations to the conservation law (8.90) and the related Hamilton-Jacobi equation

$$\partial_t v + H(\partial_x v) = 0,$$

based on viscous approximations. We will also see that too little viscosity may give unstable approximations.

To show the difficulties to solve numerically a problem like (8.90) and (8.89) we consider a related steady-state problem (i.e. a problem that has no dependence on t)

$$\begin{aligned} &\partial_x w(x) - \varepsilon \ \partial_x^2 w(x) = 0 \quad \text{for } x < 0, \\ &\lim_{x\to-\infty} w(x) = 1, \ w(0) = 0, \end{aligned} \tag{8.91}$$

where $\varepsilon \geq 0$ is fixed. It is easy to verify that the exact solution is $w(x) = 1 - \exp(\frac{x}{\varepsilon})$, for $x \leq 0$. Now, we construct a uniform partition of $(-\infty, 0]$ with nodes $x_j = j\Delta x$ for $j = 0, -1, -2, \ldots$, where $\Delta x > 0$ is a given mesh size. Denoting by W_j the approximation of $w(x_j)$, the use of a second order accurate finite element method or finite difference scheme method leads to the scheme

$$\begin{aligned} \frac{W_{j+1} - W_{j-1}}{2\Delta x} - \varepsilon \ \frac{W_{j+1} - 2W_j + W_{j-1}}{(\Delta x)^2} &= 0, \ j = -N+1, \ldots, -1, \\ W_0 &= 0, \\ W_{-N} &= 1. \end{aligned} \tag{8.92}$$

Assume that N is odd. If $\varepsilon \ll \Delta x$, the solution of (8.92) is approximated by

$$\frac{W_{j+1} - W_{j-1}}{2\Delta x} = 0,$$

which yields the oscillatory solution $W_{2i} = 0$ and $W_{2i+1} = 1$ that does not approximate w, instead $\|w - W\|_{L^2} = \mathcal{O}(1)$. One way to overcome this difficulty is to replace, in (8.92), the *physical diffusion* ε by the *artificial diffusion* $\hat{\varepsilon} = \max\{\varepsilon, \frac{\Delta x}{2}\}$. For the general problem $\beta \cdot \nabla u - \varepsilon \Delta u = f$ take $\hat{\varepsilon} = \max\{\varepsilon, |\beta|\frac{\Delta x}{2}\}$. Now, when $\varepsilon \ll \Delta x$, we have $\hat{\varepsilon} = \frac{\Delta x}{2}$ and the method (8.92), with ε replaced by $\hat{\varepsilon}$, yields $W_j = W_{j-1}$ for $j = -N+1, \ldots, -1$, that is $W_j = 1$ for $j = -N, \ldots, -1$, which is an acceptable solution with $\|w - W\|_{L^2} = \mathcal{O}(\sqrt{\Delta x})$. Another way to cure the problem is to resolve by choosing Δx small enough, so that $\hat{\varepsilon} = \varepsilon$.

The Lax-Friedrich method for the problem (8.90), is given by

$$U_j^{n+1} = U_j^n - \Delta t \left[\frac{H(U_{j+1}^n) - H(U_{j-1}^n)}{2\Delta x} - \frac{(\Delta x)^2}{2\Delta t} \ D_+ D_- U_j^n \right], \tag{8.93}$$

with

$$D_+ V_j = \frac{V_{j+1} - V_j}{\Delta x}, \ D_- V_j = \frac{V_j - V_{j-1}}{\Delta x} \text{ and } D_+ D_- V_j = \frac{V_{j+1} - 2V_j + V_{j-1}}{(\Delta x)^2}.$$

The stability condition for the method (8.93) is

$$\lambda \equiv \frac{\Delta x}{\Delta t} > \max_u |H'(u)|. \tag{8.94}$$

We want to approximate the viscosity solution of the one-dimensional Hamilton-Jacobi equation

$$\partial_t v + H(\partial_x v) = 0, \tag{8.95}$$

where $v = \lim_{\varepsilon \to 0^+} v^\varepsilon$ and

$$\partial_t v^\varepsilon + H(\partial_x v^\varepsilon) = \varepsilon \ \ \partial_x^2 v^\varepsilon. \tag{8.96}$$

Setting $u = \partial_x v$ and taking derivatives in (8.95), we obtain a conservation law for u, that is

$$\partial_t u + \partial_x H(u) = 0. \tag{8.97}$$

To solve (8.95) numerically, a basic idea is to apply (8.93) on (8.97) with $U_i^n = (V_{i+1}^n - V_{i-1}^n)/(2\Delta x)$ and then use summation over i to approximate the integration in (8.80). We get

$$\begin{aligned} \frac{V_{j+1}^{n+1} - V_{j-1}^{n+1}}{2\Delta x} &= \frac{V_{j+1}^n - V_{j-1}^n}{2\Delta x} \\ -\Delta t \Bigg[\frac{H\Big(\frac{V_{j+2}^n - V_j^n}{2\Delta x}\Big) - H\Big(\frac{V_j^n - V_{j-2}^n}{2\Delta x}\Big)}{2\Delta x} &- \frac{(\Delta x)^2}{2\Delta t} \ \ D_+ D_- \frac{V_{j+1}^n - V_{j-1}^n}{2\Delta x} \Bigg]. \end{aligned}$$

Summing over j and using that $V_{-\infty}^m = 0$ and $H(0) = 0$, it follows that

$$V_j^{n+1} = V_j^n - \Delta t \left[H\Big(\frac{V_{j+1}^n - V_{j-1}^n}{2\Delta x}\Big) - \frac{(\Delta x)^2}{2\Delta t} D_+ D_- V_j^n \right], \tag{8.98}$$

which is the Lax-Friedrich method for (8.95). Note that (8.98) is a second order accurate central difference approximation of the equation

$$\partial_t v + H(\partial_x v) = \frac{(\Delta x)^2}{2\Delta t}\big(1 - (\frac{\Delta t}{\Delta x} H')^2\big)\partial_x^2 v,$$

which is (8.96) with artificial diffusion $\Delta x(\lambda^2 - (H')^2))/(2\lambda)$.

In the two-dimensional case a first order Hamilton-Jacobi equation has the form

$$\partial_t v + H(\partial_{x_1} v, \partial_{x_2} v) = 0. \tag{8.99}$$

The analogous scheme to (8.98) for that equation is

$$\begin{aligned} V_{j,k}^{n+1} = V_{j,k}^n - \Delta t \Bigg[& H\Big(\frac{V_{j+1,k}^n - V_{j-1,k}^n}{2\Delta x_1}, \frac{V_{j,k+1}^n - V_{j,k-1}^n}{2\Delta x_2}\Big) \\ & - \frac{(\Delta x_1)^2}{4\Delta t} \frac{V_{j+1,k}^n - 2V_{j,k}^n + V_{j-1,k}^n}{(\Delta x_1)^2} \\ & - \frac{(\Delta x_2)^2}{4\Delta t} \frac{V_{j,k+1}^n - 2V_{j,k}^n + V_{j,k-1}^n}{(\Delta x_2)^2} \Bigg] \end{aligned}$$

which for $\Delta x_1 = \Delta x_2 = h$ and $\lambda = h/\Delta t$ corresponds to a second order approximation of the equation

$$\partial_t v^h + H(\partial_{x_1} v^h, \partial_{x_2} v^h) = \frac{\Delta x^2}{4\Delta t} \sum_i \partial_{x_i x_i} v - \sum_{i,j} \frac{\Delta t}{2} \partial_{x_i} H \partial_{x_j} H \partial_{x_i x_j} v.$$

Chapter 9

Rare Events and Reactions in SDE

Transition between stable equilibrium solutions are used to model for instance reaction paths and reaction rates in chemistry and nucleation phenomena in phase transitions exited by thermal fluctuations. An example of such nucleation in an under cooled liquid is the formation of the initial crystal that starts to grow to a whole solid, taking place every year in the first cold calm winter night in Swedish lakes. Deterministic differential equations cannot model such transitions between equilibrium states, since a deterministic solution never escapes from a stable equilibrium. This section shows how stochastic differential equations are used to model reaction paths and its rates, using large deviation theory from an optimal control perspective.

Let us start with a determinstic model

$$\dot{X}^t = -V'(X^t) \quad t > 0,$$

where the potential $V : \mathbb{R} \to \mathbb{R}$ is a scalar double well function, see Figure 9.2, with two stable equilibrium points x_+ and x_-, and one unstable equilbrium point x_0 in between. We see from the phase portrait Figure **??** that

$$\lim_{t\to\infty} X^t = \begin{cases} x_- & \text{if } X^0 < x_0 \\ x_+ & \text{if } X^0 > x_0 \\ x_0 & \text{if } X^0 = x_0, \end{cases} \tag{9.1}$$

which means that a path from one stable equilibrium point to another stable equilibrium point is not possible in this deterministic setting.

The stochastic setting

$$\mathrm{d}X^t = -V'(X^t)\mathrm{d}t + \sqrt{2\epsilon}\,\mathrm{d}W^t \tag{9.2}$$

can model transitions between x_- and x_+. In this section we focus on the case when the positive parameter ϵ (which measures the temperature in the chemistry model) is small, that is we study a small stochastic perturbation of the deterministic case. By

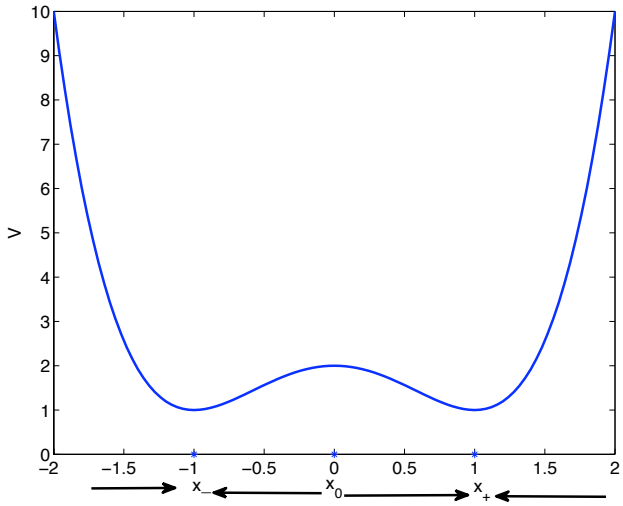


Figure 9.1: Illustration of a double well with two local minima points at x_- and x_+ and one local maximum point atx_0.

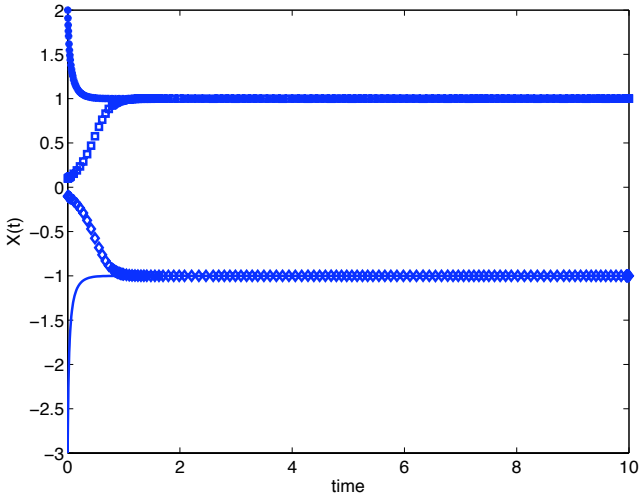


Figure 9.2: Four paths X^t from a double well potential with two local minima points at x_- and x_+ and one local maximum point at x_0.

introducing noise in the model, we may ask what is the probability to jump from one well to the other; since ϵ is small these transitions will be rare events. More precisely we shall for the model (9.2) determine:

- the invariant probability distribution and convergence towards it as time tends to infinity,
- the asymptotic behaviour of jumps from one well to another, i.e. reaction rates and reaction paths.

9.1 Invariant Measures and Ergodicity

Consider now a stochastic differential equation

$$\mathrm{d}X^t = -V'(X^t)\mathrm{d}t + \sqrt{2\epsilon}\,\mathrm{d}W^t \tag{9.3}$$

with a potential $V : \mathbb{R}^d \to \mathbb{R}$ that is smooth and satisfies $\int_{\mathbb{R}^d} e^{-V(x)/\epsilon}dx < \infty$, which implies that $V(x) \to \infty$ as $|x| \to \infty$. We also assume a global Lipschitz bound on V' to have a well defined solution X, but the global Lipschitz bound can be relaxed. The probability density for an SDE solves the Fokker-Planck equation 4.9. Sometimes this has a time independent solution - the corresponding probability measure is called an *invariant measure*. It is called invariant because if we start with this probability measure as initial probability distribution, the probability distribution obtained from the Fokker-Plank equation for later time remains unchanged, i.e. this probability distribution is time invariant. In the case of an SDE with additive noise and a drift that is the gradient of a potential function, as in (9.3), the invariant measure can be explicitly computed:

Theorem 9.1. *The SDE-model (9.3) has the invariant measure*

$$\Big(\int_{\mathbb{R}^d} e^{-V(x)/\epsilon}dx\Big)^{-1} e^{-V(x)/\epsilon}dx.$$

Proof. The Fokker-Planck equation corresponding to the dynamics (9.3) takes the form

$$\partial_t p - \partial_x\big(V'(x)p(x)\big) - \epsilon\partial_{xx}p = 0. \tag{9.4}$$

The condition to have an invariant solution means that it is time independent, i.e. $\partial_t p = 0$, and the Fokker-Planck equation can be solved explicitly

$$\epsilon p' + V'p = c,$$

for a contant c. The density p should be integrable, and consequently $p(x)$ and $p'(x)$ must tend to zero as $|x|$ tends to infinity. Therefore we have $c = 0$, which implies

$$\int \frac{dp}{p} = -\int \frac{V'}{\epsilon}dx,$$

with the solution

$$\log p(x) = C' - \frac{V(x)}{\epsilon} \quad \text{for a constant } C',$$

so that for another constant C

$$p(x) = Ce^{-V(x)/\epsilon}.$$

The requirement that $\int_{\mathbb{R}^d} p(x)dx = 1$ determines the constant to be $C = \big(\int_{\mathbb{R}^d} e^{-V(x)/\epsilon}dx\big)^{-1}$. □

□

A Monte-Carlo method to compute expected values $\int_{\mathbb{R}^d} g(y)p_0(y)dy$ in an equilibrium environment (with invariant density p_0) is typically based on approximations of the integral $T^{-1}\int_0^T g(X^t)dt$ for large T; therefore it is important to understand some basic conditions and properties of such approximations, which is the purpose if the next two theorems.

Theorem 9.2. *If one starts with any initial probability densitity and the density converges time asymptotically to the invariant density p_0, i.e. for any $\tau > 0$ the pointwise limit*

$$\lim_{t\to\infty} \tau^{-1}\int_t^{t+\tau} p^s ds = p^0$$

is satisfied, then for any continuous bounded function $g : \mathbb{R}^d \to \mathbb{R}$ there holds in the weak sense

$$\lim_{T\to\infty} T^{-1}\int_0^T g(X^t)dt = \int_{\mathbb{R}^d} g(y)p_0(y)dy. \tag{9.5}$$

We say that the stochastic process X is *ergodic* and that the invariant measure, p_0, is ergodic if (9.5) holds for all bounded continuous g.

Proof. The proof has two steps - to verify that the expected value converges and then estimate the deviation from this limit.

Step 1. By the assumption of the converging density we have

$$\begin{aligned}
\lim_{T\to\infty} \mathbb{E}[T^{-1}\int_0^T g(X^t)dt] &= \lim_{T\to\infty} \mathbb{E}\Big[T^{-1}\big(\int_0^{T^{1/2}} g(X^t)dt + \int_{T^{1/2}}^T g(X^t)dt\big)\Big] \\
&= \lim_{T\to\infty} \mathbb{E}\Big[T^{-1}\int_0^{T^{1/2}} g(X^t)dt + T^{-1}\sum_{n=T^{1/2}}^{T-1}\int_n^{n+1} g(X^t)dt\Big] \\
&= \underbrace{\int_{\mathbb{R}^d} g(y)p_0(y)dy}_{=:\mathbb{E}_0[g]},
\end{aligned}$$

where the first integral tends to zero, since g is bounded and $T^{1/2}/T \to 0$, and the $T - T^{1/2}$ integrals in the sum converge by the assumption, as explained in Example 9.4.

Step 2. Let $T = M\tau$ for some large τ, M and write the integral as a sum over M terms

$$T^{-1}\int_0^T g(X^t)dt = M^{-1}\sum_{n=1}^{M}\tau^{-1}\int_{n\tau}^{(n+1)\tau} g(X^t)dt.$$

If these terms were independent, the law of large numbers would show that the sum converges almost surely, as M tends to infinity. Since the terms are only asymptotically independent as $\tau \to \infty$, we need some other method: we shall use Chebyshevs inequality to prove convergence in probability. Let $\xi_n := \tau^{-1}\int_{n\tau}^{(n+1)\tau}\big(g(X^t) - \mathbb{E}_0[g]\big)dt$, we want to verify that for any $\gamma > 0$

$$\lim_{M,\tau\to\infty} P\Big(\frac{|\sum_{n=1}^{M}\xi_n|}{M} > \gamma\Big) = 0. \tag{9.6}$$

Chebeshevs inequality implies

$$\begin{aligned}
&P(|\sum_{n=1}^{M}\xi_n/M| > \gamma)\\
&\leq \gamma^{-2}\mathbb{E}[\sum_n\sum_m \xi_n\xi_m/M^2]\\
&= \gamma^{-2}M^{-2}\sum_n\sum_m \tau^{-2}\mathbb{E}\Big[\int_n \big(g(X^t) - \mathbb{E}_0[g]\big)dt\int_m \big(g(X^s) - \mathbb{E}_0[g]\big)ds\Big]\\
&= 2\gamma^{-2}M^{-2}\sum_{n>m}\tau^{-2}\int_n\int_m \mathbb{E}\Big[\mathbb{E}\big[\big(g(X^t) - \mathbb{E}_0[g]\big)\big(g(X^s) - \mathbb{E}_0[g]\big) \mid X^s\big]\Big]dtds\\
&\quad + \gamma^{-2}M^{-2}\sum_n\Big(\tau^{-1}\int_n \mathbb{E}\big[g(X^t) - \mathbb{E}_0[g]\big]dt\Big)^2\\
&=: I
\end{aligned}$$

and since the density p^t converges we can for each $\delta > 0$ chose τ sufficiently large so that

$$\begin{aligned}
I &= 2\gamma^{-2}M^{-2}\sum_{n>m}\tau^{-2}\int_n\int_m \mathbb{E}\Big[\int_{\mathbb{R}^d} g(y)\big(p^t(y) - p_0(y)\big)dy\big(g(X^s) - \mathbb{E}_0[g]\big)\Big]dtds\\
&\quad + \gamma^{-2}M^{-2}\sum_n\Big(\tau^{-1}\int_n \mathbb{E}[g(X^t) - \mathbb{E}_0[g]]dt\Big)^2\\
&\leq \gamma^{-2}\delta + C\gamma^{-2}M^{-1}
\end{aligned}$$

which proves (9.6). □

□

Theorem 9.3. *The process X generated by (9.3) is ergodic for positive ϵ.*

Proof. Theorem 9.2 tells us that it remains to verify that the probability density converges time asymptotically to the invariant density. Let p_0 be the invariant solution and define the *entropy*

$$E^t := \int_{\mathbb{R}^d} p \log \frac{p}{p_0} dx.$$

We know from Corollary 4.9 that p is non negative. The proof has three steps: to show that the entropy decays, that the entropy is non negative, and that the decaying entropy implies convergence of the density to the invariant density.

Step 1. Show that $\dot{E}^t = -\epsilon^{-1} \int |\epsilon p' + V'p|^2 p^{-1} dx$. Differentiation, the Fokker-Planck equation (9.4), and integration by parts[1] imply

$$\begin{aligned}
\dot{E}^t &= \int_{\mathbb{R}^d} \partial_t p \log \frac{p}{p_0} + \partial_t p \frac{p}{p} dx \\
&= \int_{\mathbb{R}^d} \underbrace{\partial_t p}_{=(V'p)'+\epsilon p''} (\log \frac{p}{p_0} + 1) dx \\
&= \int_{\mathbb{R}^d} \big((V'p)' + \epsilon p''\big)(\log \frac{p}{p_0} + 1) dx \\
&= -\int_{\mathbb{R}^d} \big(V'p + \epsilon p'\big) \cdot (\frac{p'}{p} - \underbrace{\frac{p_0'}{p_0}}_{-V'/\epsilon}) dx \\
&= -\epsilon^{-1} \int_{\mathbb{R}^d} |V'p + \epsilon p'|^2 p^{-1} dx.
\end{aligned}$$

Step 2. Show that $E^t \geq 0$ using that p and p_0 have the same mass and that $\log x$ is concave. We have

$$E^t = \int_{\mathbb{R}^d} p \log \frac{p}{p_0} dx = \int_{\mathbb{R}^d} p(-\log \frac{p_0}{p} + \frac{p_0}{p} - 1) dx$$

and the concavity of the logarithm implies $\log x \leq x - 1$, which establishes $E^t \geq 0$.

Step 3. Time integration of Step 1 gives

$$E^T + \epsilon^{-1} \int_0^T \int_{\mathbb{R}^d} |\epsilon p' + V'p|^2 p^{-1} dx dt = E^0, \tag{9.7}$$

and since E^T is non negative and E^0 is assumed to be bounded, we see that the integral $\int_0^T \int |\epsilon p' + V'p|^2 p^{-1} dx dt$ also is bounded uniformly in T. Therefore we have, for any $\tau > 0$, that $\tau^{-1} \int_t^{t+\tau} \epsilon p'^s + V'p^s ds \to 0$ in $L^2(\mathbb{R}^d)$ as $t \to \infty$, which gives $\tau^{-1} \int_t^{t+\tau} p^s ds \to p_0$ as follows: integration of

$$\epsilon p'^t + V'p^t =: f^t$$

shows that

$$p(x,t) = e^{-V(x)/\epsilon} \big(C + \int_0^x f(y,t) e^{V(y)/\epsilon} dy\big)$$

[1] A better way, in the sense of requiring less assumptions, is to directly study the Fokker-Planck equation in its weak form; then the integration by parts is not needed and (9.7) is obtained directly.

so that $\tau^{-1}\int_t^{t+\tau} p^s ds \to p_0$ as $t \to \infty$, since $\tau^{-1}\int_t^{t+\tau} f^s ds \to 0$ in $L^2(\mathbb{R}^d)$. □

□

Example 9.4 (No mass escapes to infinity)**.** The aim here is to verify that the pointwise limit $\lim_{\tau\to\infty}\int_\tau^{\tau+1} p^t dt = p_0$ implies the weak limit

$$\lim_{\tau\to\infty}\int_\tau^{\tau+1}\int_{\mathbb{R}^d} g p^t dx dt = \int_{\mathbb{R}^d} g p_0 dx, \tag{9.8}$$

for any bounded continuous function g.

Let $\bar{p}^\tau := \int_\tau^{\tau+1} p^t dt$ and define $\phi : (0,\infty) \to \mathbb{R}$ by $\phi(x) = x \log x/p_0$. The function ϕ is convex and Jensen's inequality implies together with (9.7)

$$\begin{aligned} E^0 &\geq \int_{\mathbb{R}^d}\int_\tau^{\tau+1} p^t \log\frac{p^t}{p_0} dt dx \\ &= \int_{\mathbb{R}^d}\int_\tau^{\tau+1} \phi(p^t) dt dx \\ &\geq \int_{\mathbb{R}^d} \phi\Big(\int_\tau^{\tau+1} p^t dt\Big) dx \\ &= \int_{\mathbb{R}^d} \phi(\bar{p}^\tau) dx. \end{aligned}$$

Therefore we have for any positive number n

$$\int_{\mathbb{R}^d} \bar{p}^\tau 1_{\{\bar{p}^\tau > np_0\}} dx \leq \frac{E^0}{\log n}. \tag{9.9}$$

We can split our integral into two

$$\int_{\mathbb{R}^d}\int_\tau^{\tau+1} g p^t dt dx = \int_{\mathbb{R}^d} g\bar{p}^\tau dx = \int_{\mathbb{R}^d} g\bar{p}^\tau 1_{\{\bar{p}^\tau > np_0\}} dx + \int_{\mathbb{R}^d} g\bar{p}^\tau 1_{\{\bar{p}^\tau \leq np_0\}} dx,$$

where dominated convergence yields

$$\lim_{\tau\to\infty}\int_{\mathbb{R}^d} g\bar{p}^\tau 1_{\{\bar{p}^\tau \leq np_0\}} dx = \int_{\mathbb{R}^d} g p_0 dx$$

and (9.9) shows that the other integral is negligible small

$$\int_{\mathbb{R}^d} g\bar{p}^\tau 1_{\{\bar{p}^\tau > np_0\}} dx \leq C/\log n$$

as $n \to \infty$, which proves the limit (9.8).

Exercise 9.5 (Invariant measure for Ornstein-Uhlenbeck)**.** Show that the invariant measure for the Ornstein-Uhlenbeck process is a normal distribution.

Exercise 9.6 (Vanishing noise density is not the deterministic density)**.** Prove that for a smooth function V on a bounded set A

$$\lim_{\epsilon\to 0+} \epsilon \log \int_A e^{-V(y)/\epsilon} dy = -\inf_{y\in A} V(y).$$

Such a limit was first studied by Laplace.

Exercise 9.7. Show that for a smooth function V on a bounded set A with a unique global minimum point y_+, the probabilty density $\frac{e^{-V(y)/\epsilon}}{\int_A e^{-V(y)/\epsilon} dy}$ has the limit expected value

$$\lim_{\epsilon\to 0+} \frac{\int_A e^{-V(y)/\epsilon}\phi(y)dy}{\int_A e^{-V(y)/\epsilon}dy} = \phi(y_+),$$

Compare this limit with the time-asymptotic "probability" density for the determinstic $\epsilon = 0$ case (9.1) and show they are different. What can be concluded about the limits $t \to \infty$ and $\epsilon \to 0+$ of the probability density?

Example 9.8 (Simulated Annealing)**.** The stochastic differential equation (9.3) can also be used to find minima of functions $V : \mathbb{R}^d \to \mathbb{R}^d$: we know that its invariant measure has the density $\frac{\int_A e^{-V(y)/\epsilon}\phi(y)dy}{\int_A e^{-V(y)/\epsilon}dy}$, which by Exercise 9.7 concentrates at $x \in \operatorname{argmin} V$. Therefore, by simulating the stochastic differential equation for very long time with decreasing ϵ one expect to have the path X most of the time in the global minimum; more precisely choose $\epsilon = \epsilon_1$ for $t \in [0, T_1], \ldots, \epsilon = \epsilon_n$ for $t \in [T_{n-1}, T_n]$, with $\epsilon_n \searrow 0+$ and $T_n \nearrow \infty$ as $n \to \infty$. This method is called simulated annealing and it can be proven to work for a precise choice of ϵ_n and T_n, see [**?**]. The advantage with the method is that a global minimum is found and the main question is to find a good combination of ϵ_n and T_n suitable for the particular V studied.

9.2 Reaction Rates

The invariant ergodic measure for X shows that there is a finite probability to reach all states from any point when $\epsilon > 0$, in contrast to the determinstic case $\epsilon = 0$; the invariant measure also shows that these probabilites are exponentially small, proportional to $e^{-V/\epsilon}$. It is practical to relate reaction rates to exit times from domains: define for X solving (9.3) and a given domain $A \in \mathbb{R}^d$ the exit time

$$\tau(X) = \inf\{t \ : X^t \notin A\}.$$

We want to understand the exit probability

$$P(\tau < T) = \mathbb{E}[1_{\tau<T}] =: q_\tau \quad \text{as } \epsilon \to 0+.$$

The Kolmogorov-backward equation shows that

$$\begin{aligned} \partial_t q_\tau - V' \cdot \partial_x q_\tau + \epsilon \partial_{xx} q_\tau = 0 \quad & \text{in } A \times (0, T) \\ q_\tau(x, \cdot) = 1 \quad & \text{on } \partial A \times (0, T) \\ q_\tau(\cdot, T) = 0 \quad & \text{on } A \times \{T\}. \end{aligned} \tag{9.10}$$

Remark 9.9 (A useless solution)**.** A naive try could be to remove the diffusion part $\epsilon\partial_{xx}q_\tau$ in (9.4); that leads to the hyperbolic equation

$$\begin{aligned} \partial_t q_\tau - V' \cdot \partial_x q_\tau &= 0 \quad \text{in } A \times (0,T) \\ q_\tau &= 1 \quad \text{on } \partial A \times (0,T) \\ q_\tau(\cdot, T) &= 0 \quad \text{on } A \times \{T\} \end{aligned} \tag{9.11}$$

which can be solved by the characteristics $\dot{y}^t = -V'(y^t)$:

$$\frac{d}{dt} q_\tau(y^t, t) = \partial_t q_\tau + \frac{dy^t}{dt} \cdot \partial_x q_\tau = \partial_t q_\tau - V' \cdot \partial_x q_\tau = 0.$$

Since the equilibrium points are stable, it turns out that all characteristics leave the domain on the upper part $t = T$ see Figure 9.3, where $q_\tau = 0$, so that the solution of (9.11) becomes $q_\tau = 0$, and that is a useless solution.

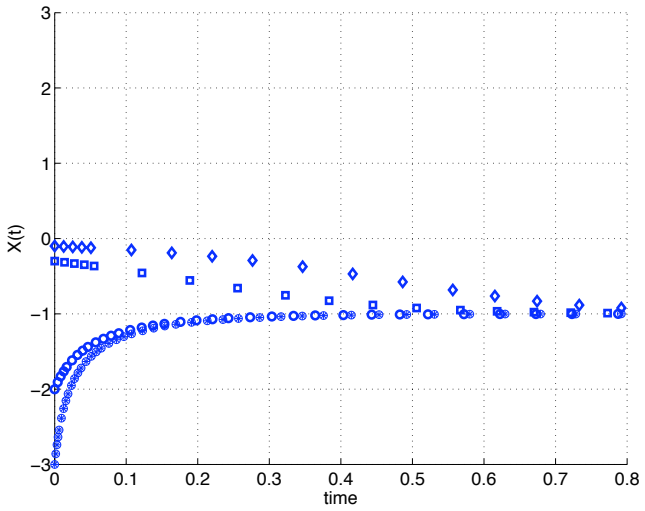


Figure 9.3: Four paths X^t starting with $X^0 < x_0$ in the domain of the global attractor x_-

The limit in Remark 9.9 needs to be refined to give something useful. The invariant measure with probabilities proportional to $e^{-V/\epsilon}$ suggests a change of variables $q_\tau(x,t) = e^{w_\epsilon(x,t)/\epsilon}$. The right way to study q_τ as $\epsilon \to 0+$ is to use the limit

$$\lim_{\epsilon \to 0+} \epsilon \log q_\tau = \lim_{\epsilon \to 0+} w_\epsilon =: w$$

which we believe has a bounded non positive limit, using the invariant measure. Since q_τ is a probability we know that $w_\epsilon \leq 0$ and (9.10) implies that w_ϵ solves the second order Hamilton-Jacobi equation

$$\begin{aligned} \partial_t w_\epsilon - V' \cdot \partial_x w_\epsilon + \partial_x w_\epsilon \cdot \partial_x w_\epsilon + \epsilon \partial_{xx} w_\epsilon &= 0 \quad \text{in } A \times (0,T) \\ w_\epsilon(x, \cdot) &= 0 \quad \text{on } \partial A \times (0,T) \\ w_\epsilon(\cdot, T) &= -\infty \quad \text{on } A \times \{T\}. \end{aligned}$$

A good way to understand this Hamilton-Jacobi equation is to view it as an optimal control problem. In the limit as ϵ tends to zero, the optimal control problem becomes determinstic, see Theorem 8.10; assume that $\lim_{\epsilon\to 0+} w_\epsilon =: w$ to obtain the first order Hamilton-Jacobi equation

$$\begin{aligned}\partial_t w \underbrace{-V'\cdot\partial_x w + \partial_x w\cdot\partial_x w}_{=:H\left(w(x),x\right)} &= 0 \quad \text{in } A\times(0,T)\\ w(x,\cdot) &= 0 \quad \text{on } \partial A\times(0,T)\\ w(\cdot,T) &= -\infty \quad \text{on } A\times\{T\}.\end{aligned}$$

Following Section 8.1.4, a useful optimal control formulation for this Hamilton-Jacobi equation is

$$\begin{aligned}\dot{Y}^t &= -V'(Y^t) + 2\alpha^t\\ \max_{\alpha:(0,T)\to\mathbb{R}^d} &-\int_0^\tau |\alpha^t|^2 dt + g(Y^\tau,\tau)\end{aligned}$$

which has the right Hamiltonian

$$\sup_{\alpha\in\mathbb{R}^d}\Big(\lambda\cdot\big(-V'(y)+2\alpha\big) - |\alpha|^2\Big) = H(\lambda,y) = -V'(y)\cdot\lambda + |\lambda|^2.$$

Here the final cost is zero, if the exit is on the boundary $\partial A\times(0,T)$, and minus infinity if the exit is on $A\times\{T\}$ (i.e. the path did not exit from A):

$$g(x,t) = \begin{cases} 0 & \text{on } \partial A\times(0,T)\\ -\infty & \text{on } A\times\{T\}.\end{cases}$$

Theorem 8.10 shows that the limit $\lim_{\epsilon\to 0+}\epsilon\log q_\tau = \lim_{\epsilon\to 0+} w_\epsilon = w$ satisfies

$$\begin{aligned}w(x,t) &= \sup_{\alpha:(t,\tau)\to\mathbb{R}^d} -\int_t^\tau |\alpha|^2 dt + g(Y^\tau,\tau)\\ &= \sup_\alpha -\frac{1}{4}\int_t^\tau |\dot{Y}^t + V'(Y^t)|^2 dt + g(Y^\tau,\tau).\end{aligned}$$

When T tends to infinity and X^0 is an equilibrium point, this limit w has a simple explicit solution showing that reaction rates are determined from local minima and saddle points of V, cf. Figure 9.4:

Theorem 9.10. *Assume that y_+ is a global attractive equilibrium in A. Let $X^0 = y_+$, then*

$$\lim_{T\to\infty}\lim_{\epsilon\to 0+}\epsilon\log q_\tau = V(y_+) - \inf_{y\in\partial A} V(y). \tag{9.12}$$

Proof. It is clear the optimal control paths starting in y_+ need to exit through ∂A, so $g(Y^\tau) = 0$. The integral cost can be rewritten as

$$\begin{aligned}&\sup_\alpha -\frac{1}{4}\int_0^\tau |\dot{Y}^t + V'(Y^t)|^2 dt \\ &= \sup_\alpha \Big(-\frac{1}{4}\int_0^\tau \underbrace{|\dot{Y}^t - V'(Y^t)|^2}_{\geq 0} dt - \underbrace{\int_0^\tau \dot{Y}^t \cdot V'(Y^t) dt}_{V(Y^\tau)-V(y_+)} \Big).\end{aligned} \tag{9.13}$$

Here the last integral is minimal if Y^τ exits through a point on ∂A where V is minimal, which is a *saddle point* if we have chose A to be the largest domain where y_+ is a global attractor. It remains to show that such an exit is compatible with having the first integral equal to zero; the first integral equals zero means that $\dot{Y}^t = V'(Y^t)$, which implies that Y *moves orthogonal to the level lines of the* V*-potential*. Such a path is possible by taking $\alpha = V'(Y^t)$ and requires T to be sufficiently large so that the time to reach the boundary on the optimal path $\dot{Y}^t = V(Y^t)$ is shorter, when X^0 tends to y_+ this time tends to infinity. □

□

We see that the probability to exit from an equilibrium is exponentially tiny, proportional to $e^{-(\inf_{y\in\partial A} V(y)-V(y_+))/\epsilon}$ as ϵ tends to zero, and therefore such exits are rare events. In the next section we show that the most probable path, the so called *reaction paths*, that gives such rare events are those where the stochastic paths X closely follow the optimal control paths Y. Since ϵ is small and the control α is not, the Brownian motion must some time be large of order $\epsilon^{-1/2}$. Therefore the rare events of exits depend on the rare events of such large deviation in the Brownian motion.

The Theorem relates to the basis of reaction theory in chemistry and statistical physics, where the probability to go from one state with energy V_1 to another with energy $V_2 > V_1$ is proportional to Boltzmanns rate $e^{-(V_2-V_1)/(k_B\mathcal{T})}$; here k_B is Boltzmanns constant and $\mathcal{T}$ is the temperature. We see that, with $\epsilon = k_B\mathcal{T}$ and V the energy, the simple model (9.3) can describe reactions and physical transition phenomena. A simple way to see that the *reaction rate* is q_τ is to take N independent particles starting in y_+. After very long time Nq_τ of them have exited from the domain and the reaction rate becomes the quotient $Nq_\tau/N = q_\tau$.

Exercise 9.11. Show that the mean exit time $u_\epsilon(x,t) := \mathbb{E}[\tau - t \mid X^t = x]$ satisfies

$$\lim_{\epsilon\to 0+} \epsilon \log u_\epsilon(y_+, t) = \inf_{y\in\partial A} V(y) - V(y_+).$$

Exercise 9.12. Does

$$\lim_{\epsilon\to 0+} \epsilon \log q_\tau = V(X^0) - \inf_{y\in\partial A} V(y)$$

hold when X^0 starts from a different point than the global attractor in A? Answer: sometimes but not in general depending on X^0 - how?

Exercise 9.11 shows that the product of the limits of the mean exit time and the probability to exit is equal to one, that is the mean exit time is exponentially large, roughly $e^{(\inf_{y\in\partial A} V(y)-V(y_+))/\epsilon}$.

9.3 Reaction Paths

This section motivates why the most probable exit paths X closely follow the optimal control paths Y. We saw in Theorem 9.10 that in the case T tending to infinity and $Y^0 = y_+$, the optimal path Y is orthogonal to the level sets of the potential V and the path starts from the minimum point y_+ (where $V(y_+) = \min_{y\in A} V(y)$) and moves towards the minimum on the boundary $\operatorname{argmin}_{y\in\partial A} V(y)$, see Figure 9.4. For bounded T the situation may change and the time to reach the boundary with the control $\alpha = V'$ may be larger than T, so that the first integral in (9.13) does not vanish and the optimal control becomes different; therefore also the exit probability is different and (9.12) is invalid; clearly such early time exit probabilities are also interesting when a rare event is unwanted, e.g. for hard-disc and power-plant failures. These most probable paths following the optimal control paths are called the *reaction paths*. Since the exit probability is small and the most probable exit path makes a *large deviation* from the equilibrium on a time span of order one, which is small compared to the expected exit time of order $e^{C/\epsilon}$ (for some positive C), the exit process can on long time spans be considered as a Poisson process with the rate $1/\mathbb{E}[\tau - t] \simeq q_\tau$.

To verify that the most probable exit paths follow the optimal control paths, we want to in some sense relate the stochastic increments $\sqrt{2\epsilon}\, dW^t$ with the control increments $\alpha^t dt$. Our first step in this direction is to find a probability measure on whole paths X, and then to see how probable the X-paths close to the optimal control paths Y_* are compared to the X-paths away from Y_*. It is clear that the probability to find $X = Y_*$ is zero, so we need to modify this argument somewhat. An informal way to understand the probability of whole paths is to consider Euler discretizations of (9.3)

$$\Big(\frac{\Delta X}{\Delta t} + V'(X_i)\Big)\Delta t = \sqrt{2\epsilon}\Delta W$$

with the probability density

$$\begin{aligned} P(\Delta W = y_i) &= e^{-\frac{|y_i|^2}{2\Delta t}} \frac{dy_i}{(2\pi\Delta t)^{d/2}} \\ &= e^{-|\frac{\Delta X}{\Delta t}+V'(X_i)|^2\Delta t/(4\epsilon)} \frac{dy_i}{(2\pi\Delta t)^{d/2}}. \end{aligned}$$

Therefore the probability measure for a whole path is

$$\begin{aligned} \Pi_{i=1}^n e^{-\frac{|y_i|^2}{2\Delta t}} \frac{dy_i}{(2\pi\Delta t)^{d/2}} &= \Pi_{i=1}^n e^{-|\frac{\Delta X}{\Delta t}+V'(X_i)|^2\Delta t/(4\epsilon)} \frac{dy_i}{(2\pi\Delta t)^{d/2}} \\ &= e^{-\sum_{i=1}^n |\frac{\Delta X}{\Delta t}+V'(X_i)|^2\Delta t/(4\epsilon)} \frac{dy_1}{(2\pi\Delta t)^{d/2}} \cdots \frac{dy_n}{(2\pi\Delta t)^{d/2}}. \end{aligned}$$

The most probable path is the one that maximises the probability density

$$e^{-\sum_{i=1}^{n} |\frac{\Delta X}{\Delta t} + V'(X_i)|^2 \Delta t/(4\epsilon)},$$

this is called the *maximum likelihood method* . In the previous section we saw that the optimal control problem does precisely this maximisation. Therefore the optimal control paths generate the most probable stochastic paths. If the density in the maximum likelihood method is almost uniform, the result is doubtful. Here the situation is the opposite - when ϵ tends to zero, the density concentrates on the most probable event, see Exercise 9.13.

If we consider W or α as perturbations, we see that the solution we have obtained is the solution of the *least-squares problems*

$$\min_W \int_0^\tau |\dot{X}^t + V'(X^t)|^2 dt = \min_\alpha \int_0^\tau |\dot{Y}^t + V'(Y^t)|^2 dt,$$

where $\dot{X}^t + V'(X^t)$ and $\dot{Y}^t + V'(Y^t)$ are the residuals, that is the error in the equation.

Exercise 9.13. In the limit as ϵ tends to zero, we saw in Exercise 9.6 that if $\int_A e^{-V(y)} dy$ is bounded, then

$$\lim_{\epsilon \to 0+} \epsilon \log \int_A e^{-V(y)/\epsilon} dy = -\inf_{y \in A} V(y).$$

Show that for a smooth function f on a bounded set A with a unique maximum point y_+, the probabilty density $\frac{e^{f(y)/\epsilon}}{\int_A e^{f(y)/\epsilon} dy}$ has the limit expected value

$$\lim_{\epsilon \to 0+} \frac{\int_A e^{f(y)/\epsilon} \phi(y) dy}{\int_A e^{f(y)/\epsilon} dy} = \phi(y_+),$$

which means that in the limit the most probable event almost surely happens and nothing else.

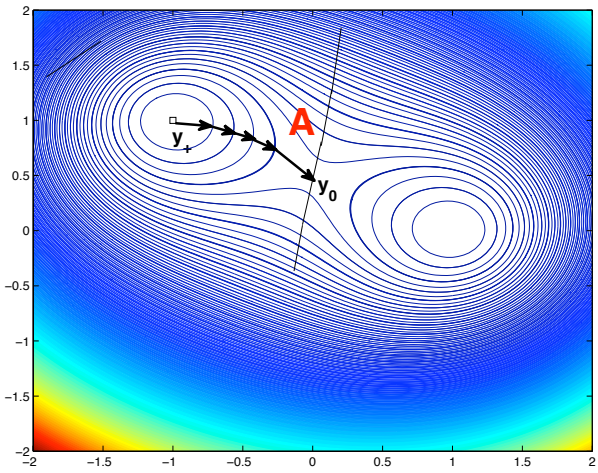


Figure 9.4: The optimal reaction path starting in the attractor y_+ moving to the sadlepoint $y_0 = \text{argmin}_{y \in \partial A}(V(y)$, inside the domain A to the left of the dashed line.

Chapter 10

Machine Learning

A central problem in machine learning is to find a neural network approximation that approximates the data well. The standard method to find such an approximation is based on a minimization problem, which is solved approximately by an iterative method called the stochastic gradient descent method. The stochastic gradient descent method is closely related numerical methods for stochastic differential equations, as we shall see in Section 10.2, in the sense that the iterations can be viewed as Euler steps of a stochastic differential equation. The convergence towards the minimum involves approximation related to time steps, as in Section 5, and time asymptotic convergence towards the equilibrium density and the rare events studied in Section 9. The approximation properties of the minimization problem is studied in Section 10.1 and the convergence of the stochastic gradient descent method in Section 10.2

10.1 Approximation with a neural network

Given a set of data points $\{(x_n, y_n) \,:\, n = 1, \ldots, N\}$, where $(x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ are independent samples from an unknown probability measure, ν, on $\mathbb{R}^d \times \mathbb{R}$, we study in this section the learning problem to determine the best neural network function $\alpha \in \mathcal{N}_K$, defined by parameters $\theta \in \mathbb{R}^{K(d+2)}$, that minimizes the expected value

$$\min_{\alpha \in \mathcal{N}_K} \mathbb{E}[g\big(y, \alpha(x, \cdot)\big)] \tag{10.1}$$

for a given loss function $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ that is convex in the second variable, α. We use the loss function

$$g\big(y, \alpha(x, \theta)\big) := |y - \alpha(x, \theta)|^2 \tag{10.2}$$

and neural network functions with one hidden layer

$$\mathcal{N}_K := \Big\{ \alpha \,:\, \alpha(x, \theta) = \begin{cases} \sum_{k=1}^K \theta_k^1 \sigma(\theta_k^2 \cdot x + \theta_k^3) =: \bar{\alpha}(x, \theta), \ x \in \mathbb{R}^d\,, & \text{for } |\theta| \le R_K \\ \big(1 - \eta(|\theta|^2)\big)\bar{\alpha}(x, \theta) + \eta(|\theta|^2)|\theta|\,, & \text{for } |\theta| > R_K \end{cases} \Big\}. \tag{10.3}$$

Here we study the activation function $\sigma : \mathbb{R} \to \mathbb{R}$ given by $\sigma(v) = \sin(v)$. Other commonly used activation functions are e.g. $\sigma(v) = 1/(1 + e^{-v})$ and $\sigma(v) = \max(v, 0)$.

The parameters $\theta = (\theta_k^1, \theta_k^2, \theta_k^3)_{k=1}^K$ are chosen as $\theta_k^1 \in \mathbb{R}$, $\theta_k^2 \in \mathbb{R}^d$, $\theta_k^3 \in \mathbb{R}$ and the non negative cut-off function $\eta \in \mathcal{C}^\infty(\mathbb{R})$, which satisfies

$$\eta(v) = \begin{cases} 0, & v < R_K^2\,, \\ 1, & v > 2R_K^2\,, \end{cases} \tag{10.4}$$

is introduced to make the loss function, g, coercive, namely

$$g\big(y, \alpha(x,\theta)\big) = \begin{cases} |y - \bar{\alpha}(x,\theta)|^2 & \text{for } |\theta| < R_K \\ |y-\theta|^2 & \text{for } |\theta| > \sqrt{2}R_K\,, \end{cases}$$

which also implies that $\nabla_\theta g\big(y, \alpha(x,\theta)\big)$ is globally Lipschitz continuous as a function of θ with a smooth transition in the range $R_K \le |\theta| \le \sqrt{2}R_K$. Here and below $|v| := (\sum_{j=1}^n v_n^2)^{1/2}$ denotes the Euclidian norm of $v = (v_1, \dots, v_n) \in \mathbb{R}^n$, for some $n \in \mathbb{N}$. The positive number R_K is chosen sufficiently large later, so that the approximation property in $\mathcal{N}_K$ is not changed substantially, see (10.16).

We may consider the data as $y_n = f(x_n) + \text{noise}$, for some function $f : \mathbb{R}^d \to \mathbb{R}$ and let

$$\mathrm{d}\bar{\nu}(\cdot) := \int_{\mathbb{R}} \mathrm{d}\nu(\cdot, y) \tag{10.5}$$

be the marginal distribution of the data $\{x_n\}_{n=1}^N$. The aim of this chapter is to study two properties of the loss function

$$G(\theta) := \int_{\mathbb{R}^d} |f(x) - \alpha(x,\theta)|^2 \mathrm{d}\bar{\nu}(x) \tag{10.6}$$

namely how well a neural network $\mathcal{N}_K$ can approximate f and how deep the local minima of the loss landscape are. We prove in Theorem 10.1 an error estimate for the minimum of G and discuss in Section 10.1.4 the depth of local minima. This depth in loss landscape, which we here call elevation gain, m, as precisely defined in (10.18), effects the convergence properties for the commonly used stochastic gradient descent method to determine neural network parameters $\bar{\theta}[n] \in \mathbb{R}^{(d+2)K}$ defined by

$$\bar{\theta}[n+1] = \bar{\theta}[n] - \nabla_\theta g\big(y_n, \alpha(x_n, \bar{\theta}[n])\big)\Delta s\,,\ n = 0, 1, 2, \dots \tag{10.7}$$

based on the learning rate $\Delta s > 0$ and a given initial random guess $\bar{\theta}[0]$. The method uses new independent samples (x_n, y_n) in each step (i.e. independent of each other) thereby forming an approximation to $\min_{\alpha \in \mathcal{N}_K} \mathbb{E}[g\big(y, \alpha(x,\cdot)\big)]$, as we shall see in Section 10.2.

10.1.1 An error estimate for neural network approximation

Assume that the Fourier transform

$$\widehat{f}(\omega) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} f(x) e^{-\mathrm{i}\omega\cdot x} \mathrm{d}x$$

of $f : \mathbb{R}^d \to \mathbb{R}$ has bounded $L^1(\mathbb{R}^d)$ norm. The classical universal approximation result by Barron in [1] proves

$$\min_{\alpha \in \mathcal{N}_K} \int_{|x| \le r} |f(x) - \alpha(x, \cdot)|^2 \mathrm{d}\bar{\nu}(x) \le \frac{(2r \| |\omega| \widehat{f}(\omega) \|_{L^1(\mathbb{R}^d)})^2}{(2\pi)^d K} ,$$

based on bounded measurable activation functions $\sigma : \mathbb{R} \to \mathbb{R}$ that satisfy $\lim_{v \to \infty} \sigma(v) = 1$ and $\lim_{v \to -\infty} \sigma(v) = 0$. We use the Monte Carlo method to show in Theorem 10.1 that this approximation rate, now based on the activation function $\sigma(v) = \sin(v)$, can be improved to

$$\min_{\alpha \in \mathcal{N}_K} \int_{\mathbb{R}^d} |f(x) - \alpha(x, \cdot)|^2 \mathrm{d}\bar{\nu}(x) \le \frac{\| \widehat{f} \|^2_{L^1(\mathbb{R}^d)}}{(2\pi)^d K} . \tag{10.8}$$

10.1.1.1 A Monte Carlo method

1. Formulation of the Monte Carlo method. The Fourier inversion theorem implies that for any $h : \mathbb{R}^d \to \mathbb{R}$ in the Schwartz class, we have

$$h(x) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \widehat{h}(\omega) e^{\mathrm{i}\omega \cdot x} \mathrm{d}\omega ,$$

where the Fourier transform $\widehat{h}(\omega) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} h(x) e^{-\mathrm{i}\omega \cdot x} \mathrm{d}x$ also is in the Schwartz class. Since $\widehat{f} \in L^1(\mathbb{R}^d)$ we have a representation of $f = \alpha \in \mathcal{N}_\infty$ based on the activation function $\sigma(y) = \sin(y)$:

$$\begin{aligned} f(x) &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \mathrm{Re}\big(\widehat{f}(\omega) e^{\mathrm{i}\omega \cdot x}\big) \mathrm{d}\omega \\ &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} |\widehat{f}(\omega)| \cos\big(\omega \cdot x + \arg \widehat{f}(\omega)\big) \mathrm{d}\omega \\ &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} |\widehat{f}(\omega)| \sin\big(\omega \cdot x + \arg \widehat{f}(\omega) + \frac{\pi}{2}\big) \mathrm{d}\omega . \end{aligned} \tag{10.9}$$

This integral can be sampled by the Monte Carlo method and the probability density function that minimizes the variance of the approximation error is $|\widehat{f}(\omega)| / \int_{\mathbb{R}^d} |\widehat{f}(\omega)| \mathrm{d}\omega$, as verified in (10.14) and Step 4. The Monte Carlo method becomes

$$\begin{aligned} \frac{1}{(2\pi)^{d/2}} \mathrm{Re} \int_{\mathbb{R}^d} \widehat{f}(\omega) e^{\mathrm{i}\omega \cdot x} \mathrm{d}\omega &\simeq \sum_{k=1}^K \frac{\mathrm{Re}\big(\widehat{f}(\bar{\omega}^k) e^{\mathrm{i}\bar{\omega}^k \cdot x}\big)}{K(2\pi)^{d/2} p(\bar{\omega}^k)} \\ &= \sum_{k=1}^K \frac{|\widehat{f}(\bar{\omega}^k)| \cos\Big(\bar{\omega}^k \cdot x + \arg \widehat{f}(\bar{\omega}^k)\Big)}{K(2\pi)^{d/2} p(\bar{\omega}^k)} \\ &=: \tilde{\alpha}(x, \bar{\omega}) \end{aligned} \tag{10.10}$$

where $\bar{\omega}^k \in \mathbb{R}^d$, $k = 1, 2, 3, \ldots K$, are independent samples with a probability density p on $\mathbb{R}^d$ and the neural network parameter in (10.3) is

$$\theta = \Big(\frac{|\widehat{f}(\bar{\omega}^k)|}{K(2\pi)^{d/2} p(\bar{\omega}^k)}, \bar{\omega}^k, \frac{\pi}{2} + \arg \widehat{f}(\bar{\omega}^k) \Big)_{k=1}^{K} . \tag{10.11}$$

We note that $\bar{\alpha}(\cdot, \theta) \in \mathcal{N}_K$ where the parameters θ_k^1 are amplitudes, θ_k^2 frequencies, and θ_k^3 phase shifts. We will use the notation $\tilde{\alpha} \in \mathcal{N}_K(p)$ for functions $\tilde{\alpha} \in \mathcal{N}_K$ defined by (10.10) where the frequencies $\bar{\omega}^k = \theta_k^2$ are sampled from the probability density $p : \mathbb{R}^d \to [0, \infty)$.

10.1.2 A property of the loss landscape

Let m be the highest elevation gain for a path from a local minimum to the global minimum in the θ loss landscape generated by the objective function $G : \mathbb{R}^d \to \mathbb{R}$, defined in (10.6). The stochastic differential equation

$$\mathrm{d}\theta_s = -\nabla G(\theta_s)\mathrm{d}s + \sqrt{2\epsilon}\, \mathrm{d}W_s \,, \quad s > 0 \,,$$

based on the standard Wiener process $W : [0, \infty) \times \Omega \to \mathbb{R}^{K(d+2)}$, is related to the stochastic gradient descent method (10.7), see [27]. Here Ω is the set of outcomes of the Wiener process and ϵ is a positive parameter. Under some assumptions on G, related to coercivity, the probability density for θ_s will, as $s \to \infty$, converge to its invariant density

$$q_\epsilon(\theta) := \frac{e^{-G(\theta)/\epsilon}}{\int_{\mathbb{R}^{(d+2)K}} e^{-G(\theta')/\epsilon} \mathrm{d}\theta'}$$

with an exponential rate $e^{-\gamma s}$, satisfying $\lim_{\epsilon \to 0+}(\epsilon \log \gamma) = -m$, see [3]. When G has a unique minimum, Laplace principle then yields $\int_{\mathbb{R}^{(d+2)K}} G(\theta) q_\epsilon(\theta) \mathrm{d}\theta \to \min_\theta G(\theta)$, as $\epsilon \to 0+$. It is therefore useful to determine the elevation gain m for neural networks. The elevation gain m for a certain neural network $\mathcal{N}_K$, with K parameters, for a supervised learning problem in dimension d is in [17] shown to decay as $m = \mathcal{O}(K^{-1/d})$.

10.1.3 Theorem 10.1: Estimation of the neural network minimum

Here follows first an elementary proof of (10.8), We use that the approximation in $\mathcal{N}_\infty$ can be characterised by probability distributions of the θ_k^2 parameters, which generate a Monte Carlo method. In the case of a finite penalty paramater, $R_K < \infty$, we use the definition

$$\begin{aligned} H(f, K, R_K) := \Big(\mathbb{E}[|f(x)|^2] + \frac{\|\widehat{f}\,\|^2_{L^1(\mathbb{R}^d)}}{(2\pi)^d K} + K\pi^2 + \frac{K \||\omega|^2 \widehat{f}\,\|_{L^1(\mathbb{R}^d)}}{\|\widehat{f}\,\|_{L^1(\mathbb{R}^d)}} \Big) \times \\ \times \frac{K \||\omega|^2 \widehat{f}\,\|_{L^1(\mathbb{R}^d)}}{\|\widehat{f}\,\|_{L^1(\mathbb{R}^d)} (R_K^2 - K\pi^2 - \frac{(\int_{\mathbb{R}^d} |\widehat{f}(\omega)| \mathrm{d}\omega)^2}{K(2\pi)^d})} \\ + \frac{K^2 \||\omega|^4 \widehat{f}\,\|_{L^1(\mathbb{R}^d)}}{\|\widehat{f}\,\|_{L^1(\mathbb{R}^d)} (R_K^2 - K\pi^2 - \frac{(\int_{\mathbb{R}^d} |\widehat{f}(\omega)| \mathrm{d}\omega)^2}{K(2\pi)^d})} \,. \end{aligned} \tag{10.12}$$

Theorem 10.1. *Assume the activation function is* $\sigma(v) = \sin(v)$ *and* $\|\widehat{f}\,\|_{L^1(\mathbb{R}^d)} < \infty$ *, then in the case* $R_K = \infty$

$$\min_{\alpha \in \mathcal{N}_K} G \leq \min_{\{p\,:\,\tilde{\alpha} \in \mathcal{N}_K(p)\}} \mathbb{E}[g\big(f(x), \alpha(x,\cdot)\big)] \leq \frac{\|\widehat{f}\,\|^2_{L^1(\mathbb{R}^d)}}{(2\pi)^d K}\,. \tag{10.13}$$

and in the case $R_K < \infty$*, if also* $\||\omega|^4 \widehat{f}\,\|_{L^1(\mathbb{R}^d)} < \infty$*,*

$$\min_{\alpha \in \mathcal{N}_K} G \leq \min_{\{p\,:\,\tilde{\alpha} \in \mathcal{N}_K(p)\}} \mathbb{E}[g\big(f(x), \alpha(x,\cdot)\big)] \leq \frac{\|\widehat{f}\,\|^2_{L^1(\mathbb{R}^d)}}{(2\pi)^d K} + H(f, K, R_K)\,,$$

where $H(f, K, R_K)$ *defined in* (10.12) *satisfies* $H(f, K, R_K) = \mathcal{O}(\frac{K^2}{R_K^2 - 20K})$.

Proof. The proof has four steps: to formulate a Monte Carlo approximation of f based on the Fourier inversion given in Section 10.1.1.1, to determine its variance, to study the effect of the penalty R_K and to determine the optimal distribution for the Monte Carlo method.

2.The variance of the Monte Carlo method. The expected squared error, with respect to the samples $\bar{\omega}^k$, that has the probability density p, and the data x, is precisely the loss function, in the case $R_K = \infty$ excluding the penalty (10.4),

$$\begin{aligned}
\mathbb{E}[g\big(f(x), \bar{\alpha}(x,\theta)\big)] &= \mathbb{E}\Big[|f(x) - \sum_{k=1}^K \frac{\mathrm{Re}\big(\widehat{f}(\bar{\omega}^k) e^{\mathrm{i}\bar{\omega}^k \cdot x}\big)}{K(2\pi)^{d/2} p(\bar{\omega}^k)}|^2\Big] \\
&= \mathbb{E}\Big[|\sum_{k=1}^K \frac{1}{K}\big(\frac{\mathrm{Re}\big(\widehat{f}(\bar{\omega}^k) e^{\mathrm{i}\bar{\omega}^k \cdot x}\big)}{(2\pi)^{d/2} p(\bar{\omega}^k)} - f(x)\big)|^2\Big] \\
&= \frac{1}{K}\mathbb{E}\Big[|\frac{\mathrm{Re}\big(\widehat{f}(\bar{\omega}^k) e^{\mathrm{i}\bar{\omega}^k \cdot x}\big)}{(2\pi)^{d/2} p(\bar{\omega}^k)} - f(x)|^2\Big] \\
&= \frac{1}{K}\Big(\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \frac{|\mathrm{Re}\big(\widehat{f}(\omega) e^{\mathrm{i}\omega \cdot x}\big)|^2}{p(\omega)} \mathrm{d}\omega - \mathbb{E}\big[|f(x)|^2\big]\Big) \\
&\leq \frac{1}{K}\Big(\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \frac{|\widehat{f}(\omega) e^{\mathrm{i}\omega \cdot x}|^2}{p(\omega)} \mathrm{d}\omega - \mathbb{E}\big[|f(x)|^2\big]\Big) \\
&= \frac{1}{K}\Big(\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \frac{|\widehat{f}(\omega)|^2}{p(\omega)} \mathrm{d}\omega - \mathbb{E}\big[|f(x)|^2\big]\Big)\,.
\end{aligned} \tag{10.14}$$

The minimum of the right hand side with respect to probability densities p is obtained for $p(\omega) = |\widehat{f}(\omega)|/\|\widehat{f}\|_{L^1(\mathbb{R}^d)}$, as verified in Step 4, which yields

$$\mathbb{E}[g\big(f(x), \bar{\alpha}(x,\theta)\big)] \leq \frac{1}{K}\Big(\frac{1}{(2\pi)^d}\big(\int_{\mathbb{R}^d} |\widehat{f}(\omega)| \mathrm{d}\omega\big)^2 - \mathbb{E}\big[|f(x)|^2\big]\Big) \tag{10.15}$$

with the convergence rate $\mathcal{O}(K^{-1})$, provided $\|\widehat{f}\|_{L^1(\mathbb{R}^d)}$ is bounded. We also have $\min_{\alpha \in \mathcal{N}_K} G \leq \mathbb{E}[g\big(f(x), \bar{\alpha}(x, \bar{\omega})\big)]$, since the minimum is less or equal to the mean and we use that $\|f\|_{L^\infty(\mathbb{R}^d)}$ is bounded since $\|\widehat{f}\,\|_{L^1(\mathbb{R}^d)}$ is bounded.

3. The case including penalty $R_K < \infty$. In the case including the penalty we have

$$f(x) - \alpha(x,\theta) = \big(1 - \eta(|\theta|^2)\big)\big(f(x) - \bar{\alpha}(x,\theta)\big) + \eta(|\theta|^2)\big(f(x) - |\theta|\big)$$

which implies

$$\begin{aligned}\mathbb{E}[g\big(f(x),\bar{\alpha}(x,\bar{\theta})\big)] &= \mathbb{E}\big[\big|\eta(|\theta|^2)\big(f(x)-|\theta|\big) + \big(1-\eta(|\theta|^2)\big)\big(f(x) - \sum_{k=1}^{K}\frac{\mathrm{Re}\big(\widehat{f}(\bar{\omega}^k)e^{\mathrm{i}\bar{\omega}^k\cdot x}\big)}{K(2\pi)^{d/2}p(\bar{\omega}^k)}\big)\big|^2\big] \\ &\leq \frac{2}{K}\Big(\big(\frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}|\widehat{f}(\omega)|\mathrm{d}\omega\big)^2 - \mathbb{E}[|f(x)|^2]\Big) + 4\mathbb{E}[(|\theta|^2+|f(x)|^2)\mathbf{1}_{|\theta|>R_K}]\,.\end{aligned}$$

Definition (10.11) implies

$$|\theta|^2 = \frac{(\int_{\mathbb{R}^d}|\widehat{f}(\omega)|\mathrm{d}\omega)^2}{K(2\pi)^d} + \sum_{k=1}^{K}|\bar{\omega}^k|^2 + \underbrace{\sum_{k=1}^{K}\big(\frac{\pi}{2} + \arg\widehat{f}(\bar{\omega}^k)\big)^2}_{=:\gamma_K}$$

therefore

$$\begin{aligned}\mathbb{E}[(|\theta|^2+|f(x)|^2)\mathbf{1}_{|\theta|>R_K}] &= \mathbb{E}[(|\theta|^2+|f(x)|^2)\mathbf{1}_{\sum_{k=1}^{K}|\bar{\omega}^k|^2 > R_K^2 - \gamma_K - \frac{(\int_{\mathbb{R}^d}|\widehat{f}(\omega)|\mathrm{d}\omega)^2}{K(2\pi)^d}}] \\ &= \mathbb{E}[(|\theta|^2+|f(x)|^2)\mathbf{1}_{\frac{\sum_{k=1}^{K}|\bar{\omega}^k|^2}{R_K^2-\gamma_K-\frac{(\int_{\mathbb{R}^d}|\widehat{f}(\omega)|\mathrm{d}\omega)^2}{K(2\pi)^d}}>1}] \\ &\leq \mathbb{E}[(|\theta|^2+|f(x)|^2)\frac{\sum_{k=1}^{K}|\bar{\omega}^k|^2}{R_K^2 - K\pi^2 - \frac{(\int_{\mathbb{R}^d}|\widehat{f}(\omega)|\mathrm{d}\omega)^2}{K(2\pi)^d}}]\end{aligned}$$

which can be written

$$\begin{aligned}&\mathbb{E}[(|\theta|^2+|f(x)|^2)\mathbf{1}_{|\theta|>R_K}] \\ &\leq \mathbb{E}[\big(|f(x)|^2 + \frac{\|\widehat{f}\,\|^2_{L^1(\mathbb{R}^d)}}{(2\pi)^d K} + K\pi^2 + \sum_{k=1}^{K}|\bar{\omega}^k|^2\big)\frac{\sum_{k=1}^{K}|\bar{\omega}^k|^2}{R_K^2 - K\pi^2 - \frac{(\int_{\mathbb{R}^d}|\widehat{f}(\omega)|\mathrm{d}\omega)^2}{K(2\pi)^d}}] \\ &= \Big(\mathbb{E}[|f(x)|^2] + \frac{\|\widehat{f}\,\|^2_{L^1(\mathbb{R}^d)}}{(2\pi)^d K} + K\pi^2 + \frac{K\||\omega|^2\widehat{f}\,\|_{L^1(\mathbb{R}^d)}}{\|\widehat{f}\,\|_{L^1(\mathbb{R}^d)}}\Big)\frac{K\||\omega|^2\widehat{f}\,\|_{L^1(\mathbb{R}^d)}}{\|\widehat{f}\,\|_{L^1(\mathbb{R}^d)}(R_K^2 - K\pi^2 - \frac{(\int_{\mathbb{R}^d}|\widehat{f}(\omega)|\mathrm{d}\omega)^2}{K(2\pi)^d})} \\ &\quad + \frac{K\||\omega|^4\widehat{f}\,\|_{L^1(\mathbb{R}^d)}}{\|\widehat{f}\,\|_{L^1(\mathbb{R}^d)}(R_K^2 - K\pi^2 - \frac{(\int_{\mathbb{R}^d}|\widehat{f}(\omega)|\mathrm{d}\omega)^2}{K(2\pi)^d})} \\ &= \mathcal{O}(\frac{K^2\mathbb{E}[|\omega|^2]^2 + K\mathbb{E}[|\omega|^4]}{R_K^2 - 20K})\,.\end{aligned}$$

We see that the penalty yields negligible contribution to the approximation error provided R_K is chosen sufficiently large, namely so that

$$\frac{K^2\mathbb{E}[|\omega|^2]^2 + K\mathbb{E}[|\omega|^4]}{R_K^2 - 20K} = o(K^{-1})\,. \tag{10.16}$$

Step 4. Optimal Monte Carlo sampling. The optimal density p is determined by minimizing the variance in (10.14)

$$\mathbb{E}[g\big(f(x),\bar{\alpha}(x,\theta)\big)] \leq \frac{1}{K}\Big(\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\frac{|\widehat{f}(\omega)|^2}{p(\omega)}\mathrm{d}\omega - \mathbb{E}\big[|f(x)|^2\big]\Big)$$

under the constraint

$$\int_{\mathbb{R}^d} p(\omega)\mathrm{d}\omega = 1\,. \tag{10.17}$$

The change of variables $p(\omega) = q(\omega)/\int_{\mathbb{R}^d} q(\omega)\mathrm{d}\omega$ implies (10.17) for any $q:\mathbb{R}^d\to[0,\infty)$. Define for any $v:\mathbb{R}^d\to\mathbb{R}$

$$H(\epsilon) := \int_{\mathbb{R}^d}\frac{|\widehat{f}(\omega)|^2}{q(\omega)+\epsilon v(\omega)}\mathrm{d}\omega\int_{\mathbb{R}^d} q(\omega)+\epsilon v(\omega)\mathrm{d}\omega\,.$$

At the optimum we have

$$\begin{aligned}0 = H'(0) &= \int_{\mathbb{R}^d}\frac{|\widehat{f}(\omega)|^2 v(\omega)}{-q^2(\omega)}\mathrm{d}\omega\underbrace{\int_{\mathbb{R}^d} q(\omega')\mathrm{d}\omega'}_{=:c_1} + \underbrace{\int_{\mathbb{R}^d}\frac{|\widehat{f}(\omega')|^2}{q(\omega')}\mathrm{d}\omega'}_{=:c_2}\int_{\mathbb{R}^d} v(\omega)\mathrm{d}\omega\\ &= \int_{\mathbb{R}^d}\big(c_2 - c_1\frac{|\widehat{f}(\omega)|^2}{q^2(\omega)}\big)v(\omega)\mathrm{d}\omega\end{aligned}$$

which implies $q(\omega) = \frac{c_1}{c_2}|\widehat{f}(\omega)|$ and consequently the optimal density becomes

$$p(\omega) = \frac{|\widehat{f}(\omega)|}{\int_{\mathbb{R}^d}|\widehat{f}(\omega')|\mathrm{d}\omega'}\,.$$

□

10.1.4 Properties of the loss landscape

Here we study a property of the loss landscape G and we let the maximum elevation gain in the deepest valley for any continuous path $\gamma_k:[0,1]\to\mathbb{R}^{(d+2)K}$ from a local minimum point θ_k to the global minimum point θ^* be denoted by

$$m := \sup_{\theta_k}\inf_{\{\gamma_k:\gamma_k(0)=\theta_k\,\&\,\gamma_k(1)=\theta^*\}}\max_{s\in[0,1]}\big(G(\gamma_k(s)) - G(\theta_k)\big) =: G(\theta_{12}) - G(\theta_1) \tag{10.18}$$

where $G(\theta_{12})$ is the highest saddle point and $G(\theta_1)$ the local minimum (i.e. not the global minimum) in the deepest valley, which is assumed to be unique.

10.2 The stochastic gradient Langevin method

The purpose of this section is to show that the learning problem (10.1) for the parameter $\theta = (\theta_k^1, \theta_k^2, \theta_k^3)_{k=1}^K \in \mathbb{R}^{K(d+2)}$ can be approximated by the stochastic gradient Langevin method

$$\bar{\theta}_{n+1} = \bar{\theta}_n - \nabla_\theta g\big(y_n, \alpha(x_n, \bar{\theta}_n)\big)\Delta s + \sqrt{2\epsilon}\Delta W_n\,,\ n = 0, 1, 2, \dots \tag{10.19}$$

based on the standard Wiener process $W : [0, \infty) \times \Omega \to \mathbb{R}^{K(d+2)}$ with increments $\Delta W_n := W\big((n+1)\Delta s\big) - W(n\Delta s)$ and time steps (i.e. learning rate) $\Delta s > 0$. The method uses new independent samples (x_n, y_n) in each step (i.e. independent of each other and of W) and we will verify that these iterations yields an approximation to $\min_{\alpha \in \mathcal{N}_K} \mathbb{E}[g\big(y, \alpha(x, \cdot)\big)]$ as $n \to \infty$ and $\epsilon \to 0+$. The stochastic gradient Langevin method is the stochastic gradient descent with additional noise, corresponding to $\epsilon > 0$ and for the case $\epsilon = 0$ it is the stochastic gradient descent method (10.7). Here Ω is the set of outcomes of the Wiener process. More precisely we show in this section that the stochastic gradient Langevin method is consistent in the sense that we can approximate the expected value (10.1) with an error that tends to zero as the learning rate Δs tends to zero and the number of neural network nodes K tends to infinity, provided the diffusion is chosen as $\epsilon \simeq K^{-1}$ and the simulation time is large enough $M\Delta s \gtrsim K^{-2}$. As an introductory step we formulate the following simple example.

Exercise 10.2 (A simple model)**.** The following example of a minimization problem to find a parameter $\theta \in \mathbb{R}$ uses similar similar iterations as (10.7). Assume we seek the minimal expected value

$$\min_{\theta\in\mathbb{R}} \mathbb{E}[f(\theta, Y)] \tag{10.20}$$

where Y is the stochastic variable with standard normal distribution (with mean zero and variance one) and $f(\theta, y) := |\theta - y|^2$ for $\theta \in \mathbb{R}$ and $y \in \mathbb{R}$.

a. Consider the iterations

$$\begin{aligned} \theta_0 &= 1 \\ \theta_{n+1} &= \theta_n - \Delta t \frac{\partial f}{\partial \theta}(\theta_n, Y_n)\,, \ n = 0, 1, 2, \dots\,, \end{aligned} \tag{10.21}$$

and show that there is a constant C such that $\mathbb{E}[|\theta_n|^2] \le C$, $n = 0, 1, 2, \dots$ for a suitable choice of $\Delta t \in \mathbb{R}$. Determine these Δt. Here, Y_n, $n = 0, 1, 2, \dots$ are independent stochastic variables which all are standard normal distributed and $\mathbb{E}[|\theta_n|^2]$ denotes the expected value of $|\theta_n|^2$. Determine also θ_* where

$$\mathbb{E}[f(\theta_*, Y)] = \min_{\theta\in\mathbb{R}} \mathbb{E}[f(\theta, Y)]$$

and establish the convergence rate of $\lim_{n\to\infty} \mathbb{E}[|\theta_n - \theta_*|^2]$.

b. Write a program that performs the iterations in problem 1a and plots θ_n, $n = 0, 1, 2, \dots N$.

c. Show that the stochastic gradient descent iterations (10.21) are in fact forward Euler steps for an Ornstein-Uhlenbeck process. Formulate also the gradient descent method, based on computing the expected value exactly, to solve the minimization problem (10.20), determine its convergence rate and compare with the result in problem 1a.

10.2.1 Convergence of the stochastic gradient Langevin method

The convergence proof of stochastic gradient descent is based on consistency analysis of the discrete stochastic gradient Langevin method (10.19), with respect to the Kolmogorov backward equation for the process $\theta : [0, \infty) \times \Omega \to \mathbb{R}^{K(d+2)}$ defined by the gradient Langevin dynamics

$$\mathrm{d}\theta_s = -\nabla_\theta \mathbb{E}[g\big(y, \alpha(x, \theta_s)\big) \mid \theta_s]\mathrm{d}s + \sqrt{2\epsilon}\,\mathrm{d}W_s\,, \tag{10.22}$$

where the expected value is with respect to the data (x, y) only. Here we use the same initial data $\theta_0 = \bar{\theta}_0$ and the same Wiener process as in (10.19). We introduce, for later use, the notation $G : \mathbb{R}^{K(d+2)} \to \mathbb{R}$ for this expected value

$$G(\theta) := \mathbb{E}[g\big(y, \alpha(x, \theta)\big) \mid \theta] = \int_{\mathbb{R}^d\times\mathbb{R}} g\big(y, \alpha(x, \theta)\big)\mathrm{d}\nu(x, y)\,.$$

The following three steps are the basic ideas to obtain the convergence rate and prove Theorem 10.3.

(a) We observe that the parameter γ, for the convergence rate $e^{-\epsilon\gamma s}$ towards the invariant measure for process θ, is related to the loss landscape of the neural network approximation: large deviation theory, related to Theorem 9.10 and further described in [18], shows that $\lim_{\epsilon\to 0+} \epsilon \log \gamma = -m$, where m is the highest elevation gain for a path from a local minimum to the global minimum in the θ loss landscape generated by the objective function $\mathbb{E}[g\big(y, \alpha(x,\theta)\big)]$.

(b) The elevation gain m for a certain neural network $\mathcal{N}_K$, with K parameters, for a supervised learning problem in dimension d is in [17] shown to decay as $m = \mathcal{O}(K^{-1/d})$.

$$\min_{\alpha\in\mathcal{N}_K} \mathbb{E}[g\big(f(x), \alpha(x,\cdot)\big)] - \min_{\alpha\in\mathcal{N}_\infty} \mathbb{E}[g\big(f(x), \alpha(x,\cdot)\big)] = \mathcal{O}(K^{-1})$$

holds when the Fourier transform of $f : \mathbb{R}^d \to \mathbb{R}$ is bounded in $L^1(\mathbb{R}^d)$. The set $\mathcal{N}_\infty$ denotes the network functions in (10.3) given by a combination of possibly infinitely many translations and dilatations of the activation function σ.

We have the following consistency result

Theorem 10.3 (Stochastic gradient Langevin convergence rate)**.** *Assume the activation function is $\sigma(y) = \sin y$, the initial parameters $|\bar\theta_0| \le R_K$ are bounded, and the data is given by $y = f(x)$, where $f : \mathbb{R}^d \to \mathbb{R}$ and its Fourier transform $\widehat{f}$ has bounded $\|(1+|\omega|^4)\widehat{f}(\omega)\|_{L^1(\mathbb{R}^d)}$ norm, the function $G : \mathbb{R}^{K(d+2)} \to \mathbb{R}$ has non degenerate unique local minima and saddle points, then there are positive constants C_τ, c, $\bar c$ and C such that for sufficiently many $M = \tau/\Delta s$ steps with the stochastic gradient Langevin method (10.19) the expected error satisfies*

$$\big|\mathbb{E}[g\big(f(x), \alpha(x,\bar\theta_M)\big)] - \min_{\alpha\in\mathcal{N}_\infty} \mathbb{E}[g\big(f(x), \alpha(x,\cdot)\big)]\big| \le C_\tau \Delta s + C(\epsilon + K^{-1} + e^{-c\epsilon\gamma\tau})\,, \quad (10.23)$$

where the expected values are over the data x, the Wiener process W and the initial data of the parameters $\bar\theta_0$ for the neural network $\mathcal{N}_K$. The constants depend on $\|\widehat{\alpha(\cdot,\bar\theta_M)}\|_{L^1(\mathbb{R}^d)}$.

We note that the constant C_τ depends on the simulation time τ. In these notes we can only conclude an exponential bound of this constant with respect to τ. It remains a challenge to obtain sharper estimates of C_τ and better estimates of γ, so that the result could be practically useful.

The convergence proof has four main steps:

(1) to verify that the stochastic gradient Langevin method samples $\nabla_\theta G(\bar\theta_n)$, i.e. the expected value of the gradient, asymptotically as $n \to \infty$, by estimating for $M = \tau/\Delta s$

$$\big|\mathbb{E}[g\big(y, \alpha(x,\bar\theta_M)\big) - g\big(y, \alpha(x,\theta_{M\Delta s})\big)]\big| \le C_\tau \Delta s\,,$$

a poor estimate $C_\tau < e^{c\tau}$ is provided in Section 10.2.5 for completeness,

(2) to first show that for any ϵ the expected value of $G(\theta_\tau)$ is well approximated by the corresponding expected value based on the Gibbs probability density

$$\mathrm{d}\mu := e^{-G(\theta)/\epsilon}\mathrm{d}\theta / \int_{\mathbb{R}^{K(d+2)}} e^{-G(\theta)/\epsilon}\mathrm{d}\theta\,,$$

using geometric ergodicity provided by Gibbs-weighted energy estimates, i.e.

$$\mathbb{E}[g\big(y,\alpha(x,\theta_{M\Delta s})\big)] - \int_{\mathbb{R}^{K(d+2)}} G(\theta)\mathrm{d}\mu(\theta) = \mathcal{O}(e^{-\epsilon\gamma\tau+c/\epsilon})\,,$$

which is a refinement of the ergodic result of Theorem 9.3 in the sense that it provides a convergence rate, and then prove that the Gibbs expected value tends to the minimum, $\min_{\alpha\in\mathcal{N}_K} G$, i.e.

$$\int_{\mathbb{R}^{K(d+2)}} G(\theta)\mathrm{d}\mu(\theta) - \min_{\alpha\in\mathcal{N}_K} G = \mathcal{O}(\epsilon)\,,$$

as the temperature parameter $\epsilon \to 0+$;

(3) to estimate the neural network minimum:

$$\min_{\alpha\in\mathcal{N}_K} G - \min_{\alpha_\in\mathcal{N}_\infty} G = \mathcal{O}(K^{-1})\,.$$

The sum of the estimates in steps (1)-(3) yields the conclusion (10.23) in Theorem 10.3, while Theorem 10.1 considers step (3). Step (1) and (2) are presented in the following sections.

The analysis here assumes that the function G has non degenerate local minima and saddle points, in order to use the large deviation theory. It would be interesting to avoid this non degeneracy assumption, since experiments show that the Hessian of the loss function is often degenerate, see [34].

We will first analyze the stochastic gradient Langevin method. Define for $s \le \tau$ the expected value of the approximate cost function

$$v(\theta', s) := \mathbb{E}[g\big(y,\alpha(x,\theta_\tau)\big)\,|\,\theta_s = \theta']\,, \tag{10.24}$$

where the expected value is with respect to both the data (x, y) and the Wiener process W in (10.19) and the process $\theta : [0,\infty)\times\Omega \to \mathbb{R}^{K(d+2)}$ solves the gradient descent Ito differential equation (10.22). The extension of the discrete stochastic gradient Langevin method to continuous time given by

$$\bar\theta_s = \bar\theta_n + (s - n\Delta s)\nabla_\theta g\big(y_n,\alpha(x_n,\bar\theta_n)\big) + \sqrt{2\epsilon}(W_s - W_{n\Delta s})\,,\ n\Delta s \le s < (n+1)\Delta s,$$

implies that it solves the Ito differential equation with piecewise constant drift

$$\mathrm{d}\bar\theta_s = -\nabla_\theta g\big(y_n,\alpha(x_n,\bar\theta_n)\big)\mathrm{d}s + \sqrt{2\epsilon}\,\mathrm{d}W_s\,,\quad n\Delta s \le s < (n+1)\Delta s\,. \tag{10.25}$$

These definitions establishes by Ito's formula and the same initial condition $\theta_0 = \bar{\theta}_0$ the error representation

$$\begin{aligned}
&\mathbb{E}[g\big(y, \alpha(x, \bar{\theta}_\tau)\big)] - \mathbb{E}[g\big(y, \alpha(x, \theta_\tau)\big)] = \mathbb{E}[v(\bar{\theta}_\tau, \tau) - v(\bar{\theta}_0, 0)] \\
&= \mathbb{E}[\int_0^\tau \mathrm{d}v(\bar{\theta}_s, s)] \\
&= \mathbb{E}[\int_0^\tau \big(\partial_s v(\bar{\theta}_s, s) - \nabla_\theta g(y_{n_s}, \alpha(x_{n_s}, \bar{\theta}_{n_s})) \cdot \nabla v(\bar{\theta}_s, s) + \epsilon \Delta v(\bar{\theta}_s, s)\big)\mathrm{d}s] \\
&\quad + \underbrace{\mathbb{E}[\int_0^\tau \sqrt{2\epsilon} \nabla v(\bar{\theta}_s, s) \cdot \mathrm{d}W_s]}_{=0},
\end{aligned}$$

where for any $s \in [0, \infty)$ the notation $n_s \in \mathbb{N}$ is defined by $n_s \Delta s \leq s < (n_s + 1)\Delta s$. The function v satisfies the Kolmogorov backward equation

$$\begin{aligned}
\partial_s v(\theta, s) &= \nabla_\theta G(\theta) \cdot \nabla v(\theta, s) - \epsilon \Delta v(\theta, s)\,, \;\; s < \tau\,, \;\; \theta \in \mathbb{R}^{K_d}\,, \\
v(\theta, \tau) &= G(\theta)\,, \;\; \theta \in \mathbb{R}^{K_d}\,,
\end{aligned}$$

where we use the notation $K_d := K(d + 2)$. We can write

$$\bar{\theta}_s = \bar{\theta}_n - \nabla_\theta g\big(y_n, \alpha(x_n, \bar{\theta}_n)\big)(s - n\Delta s) + \sqrt{2\epsilon}(W_s - W_{n\Delta s}) =: \bar{\theta}_n + \Delta\xi_s$$

which implies

$$\begin{aligned}
&\mathbb{E}[g\big(y, \alpha(x, \bar{\theta}_\tau)\big)] - \mathbb{E}[g\big(y, \alpha(x, \theta_\tau)\big)] \\
&= \mathbb{E}\Big[\int_0^\tau \nabla v(\bar{\theta}_s, s) \cdot \Big(\nabla_\theta G(\bar{\theta}_s) - \nabla_\theta g\big(y_{n_s}, \alpha(x_{n_s}, \bar{\theta}_{n_s})\big)\Big)\mathrm{d}s\Big] \\
&= \int_0^\tau \mathbb{E}[\big(\nabla v(\bar{\theta}_s, s) - \nabla v(\bar{\theta}_{n_s}, s)\big) \cdot \Big(\nabla_\theta G(\bar{\theta}_s) - \nabla_\theta g\big(y_{n_s}, \alpha(x_{n_s}, \bar{\theta}_{n_s})\big)\Big)]\mathrm{d}s \\
&\quad + \int_0^\tau \mathbb{E}[\nabla v(\bar{\theta}_{n_s}, s) \cdot \Big(\nabla_\theta G(\bar{\theta}_{n_s} + \Delta\xi_s) - \nabla_\theta g\big(y_{n_s}, \alpha(x_{n_s}, \bar{\theta}_{n_s})\big)\Big)]\mathrm{d}s\,.
\end{aligned} \tag{10.26}$$

By using the same technique as to prove weak convergence of Eulers method in Theorem 5.8, the last term in (10.26) is by the a priori estimates in Lemma 10.4 in Section 10.2.5 of order one in Δs,

$$|\int_0^\tau \mathbb{E}[\nabla v(\bar{\theta}_{n_s}, s) \cdot \Big(\nabla G(\bar{\theta}_{n_s} + \Delta\xi_s) - \nabla g(y_{n_s}, \alpha(x_{n_s}, \bar{\theta}_{n_s}))\Big)]\mathrm{d}s| \leq C_\tau \Delta s, \tag{10.27}$$

since the integrand is zero for $s = n\Delta s$ and it has a bounded derivative with respect to s in the interval $n\Delta s \leq s < (n + 1)\Delta s$, using the expected value to cancel the $\mathrm{d}W$ contribution from the time derivative, as shown in detail by Lemma 10.5 in Section 10.2.6. Similarly the integrand

$$\mathbb{E}[\big(\nabla v(\bar{\theta}_s, s) - \nabla v(\bar{\theta}_{n_s}, s)\big) \cdot \Big(\nabla G(\bar{\theta}_s) - \nabla_\theta g\big(y_{n_s}, \alpha(x_{n_s}, \bar{\theta}_{n_s})\big)\Big)]$$

also vanishes for $s = n\Delta s$ and has a bounded derivative with respect to s. Therefore, the first term in the right hand side of (10.26) is also bounded by $\mathcal{O}(\Delta s)$, provided the second derivatives of g and v and the third derivative of ϵv with respect to θ are bounded. The poor exponential bound $C_\tau \leq e^{c\tau}$ provided in Section 10.2.6 excludes more precise computational complexity studies with respect to the number of stochastic gradient descent steps M.

10.2.2 Convergence to the minimum

This section shows that

$$|\mathbb{E}[g\big(y, \alpha(x, \theta_\tau)\big)] - \min_{\theta \in \mathbb{R}^{K_d}} G(\theta)| = \mathcal{O}(\epsilon + e^{-c'\tau + c''/\epsilon}) \,,$$

for positive constant c', c''. The study is split into two steps. The first step, presented in Subsection 10.2.3, relates the Gibbs measure

$$\frac{e^{-G(\theta)/\epsilon}}{\int_{\mathbb{R}^{K_d}} e^{-G(\theta')/\epsilon} \mathrm{d}\theta'} \,,$$

which is the invariant probability density for the process (10.22), to its minimum by deriving

$$\int_{\mathbb{R}^{K_d}} G(\theta) \frac{e^{-G(\theta)/\epsilon}}{\int_{\mathbb{R}^{K_d}} e^{-G(\theta')/\epsilon} \mathrm{d}\theta'} \mathrm{d}\theta - \min_{\theta \in \mathbb{R}^{K_d}} G(\theta) = \mathcal{O}(\epsilon) \,.$$

The second step, presented in Subsection 10.2.4, shows that the process θ_τ is geometrically ergodic by proving

$$\Big|\mathbb{E}[g\big(y, \alpha(x, \theta_\tau)\big)] - \int_{\mathbb{R}^{K_d}} G(\theta) \frac{e^{-G(\theta)/\epsilon}}{\int_{\mathbb{R}^{K_d}} e^{-G(\theta')/\epsilon} \mathrm{d}\theta'} \mathrm{d}\theta\Big| = \mathcal{O}(e^{-\epsilon\gamma\tau/2 + c/\epsilon})$$

where γ and c are positive constants.

10.2.3 The Gibbs measure

To analyze the invariant probability density for the Kolmogorov forward equation applied to the process (10.22) we introduce the notation $G(\theta^*) := \min_{\theta \in \mathbb{R}^{K_d}} G(\theta)$ and $\mathcal{E}(\theta) := G(\theta) - G(\theta^*)$. Dominated convergence implies

$$\lim_{\epsilon \to 0+} \int_{\mathbb{R}^{K_d}} G(\theta) \frac{e^{-G(\theta)/\epsilon}}{\int_{\mathbb{R}^{K_d}} e^{-G(\theta')/\epsilon} \mathrm{d}\theta'} \mathrm{d}\theta = G(\theta^*) + \lim_{\epsilon \to 0+} \int_{\mathbb{R}^{K_d}} \mathcal{E}(\theta) \frac{e^{-\mathcal{E}(\theta)/\epsilon}}{\int_{\mathbb{R}^{K_d}} e^{-\mathcal{E}(\theta')/\epsilon} \mathrm{d}\theta'} \mathrm{d}\theta = G(\theta^*) \,.$$

To prove the convergence rate

$$\int_{\mathbb{R}^{K_d}} \mathcal{E}(\theta) \frac{e^{-\mathcal{E}(\theta)/\epsilon}}{\int_{\mathbb{R}^{K_d}} e^{-\mathcal{E}(\theta')/\epsilon} \mathrm{d}\theta'} \mathrm{d}\theta = \mathcal{O}(\epsilon) \tag{10.28}$$

we write $\mathcal{E}(\theta) = (\theta - \theta^*) \cdot \bar{G}''(\theta - \theta^*, \theta^*)(\theta - \theta^*)$ where

$$\bar{G}''(\theta - \theta^*, \theta^*) := \int_0^1 \int_0^1 G''\big(\theta^* + st(\theta - \theta^*)\big) t \mathrm{d}t \mathrm{d}s \,.$$

The change of variables $\theta' := (\theta - \theta^*)/\sqrt{\epsilon}$ implies

$$\int_{\mathbb{R}^{K_d}} \mathcal{E}(\theta) \frac{e^{-\mathcal{E}(\theta)/\epsilon}}{\int_{\mathbb{R}^{K_d}} e^{-\mathcal{E}(\theta')/\epsilon} \mathrm{d}\theta'} \mathrm{d}\theta = \epsilon \, \frac{\int_{\mathbb{R}^{K_d}} \theta' \cdot \bar{G}''(\epsilon^{1/2}\theta', \theta^*)\theta' e^{-\theta' \cdot \bar{G}''(\epsilon^{1/2}\theta', \theta^*)\theta'} \mathrm{d}\theta'}{\int_{\mathbb{R}^{K_d}} e^{-\theta' \cdot \bar{G}''(\epsilon^{1/2}\theta', \theta^*)\theta'} \mathrm{d}\theta'} = \mathcal{O}(\epsilon) \,, \tag{10.29}$$

since the combination of dominated convergence and the assumption that the Hessian $G''(\theta^*)$ is positive definite (uniformly in K) show that the integrals in the right hand side have finite positive limits as $\epsilon \to 0+$.

10.2.4 Geometric ergodicity

The approximate minimization property (10.28) of the stochastic gradient Langevin method obtained for small ϵ is important for our error analysis of the learning problem (10.1). In practice, we cannot use a limit parameter $\lim_{\tau \to \infty} \theta_\tau =: \theta_\infty$ (or $\bar{\theta}_\infty$) since we only have a finite number of stochastic gradient descent iterations. To study the asymptotics of the dynamics (10.19) as $\tau \to \infty$, we will present well known weighted energy estimates based on the Kolmogorov backward equation for the expected value $\hat{v}(\theta, t) := \mathbb{E}[g\big(y, \alpha(x, \theta_t)\big) \,|\, \theta_0 = \theta]$, obtained by letting $t = \tau - s$ in (10.24),

$$\begin{aligned} \partial_t \hat{v}(\theta, t) + \nabla G(\theta) \cdot \nabla \hat{v}(\theta, t) - \epsilon \Delta \hat{v}(\theta, t) = 0 \,, \quad t > 0 \,, \ \theta \in \mathbb{R}^{K_d} \\ \hat{v}(\theta, 0) = G(\theta) \,, \ \theta \in \mathbb{R}^{K_d} \,. \end{aligned} \tag{10.30}$$

The same partial differential equation holds for $\bar{v} = \hat{v} - c$, where c is any constant. Assume that the conditions (10.39) for G holds, then the equilibrium measure $\mathrm{d}\mu := e^{-G/\epsilon}\mathrm{d}\theta / \int_{\mathbb{R}^{K_d}} e^{-G/\epsilon}\mathrm{d}\theta$ exists. We have

$$e^{G/\epsilon} \nabla \cdot (e^{-G/\epsilon} \epsilon \nabla v) = -\nabla G \cdot \nabla v + \epsilon \Delta v := \mathcal{L} v \,.$$

Therefore the operator $\mathcal{L}$ is symmetric in the weighted Sobolev space $H^1(\mathrm{d}\mu)$ and we obtain the variational equation: find $\bar{v} \in L^2(\mathbb{R}_+; H^1(\mathrm{d}\mu))$ with $\partial_t \bar{v} \in L^2(\mathbb{R}_+; H^{-1}(\mathrm{d}\mu))$ such that

$$\int_{\mathbb{R}^{K_d}} \partial_t \bar{v}(\theta, t) w(\theta) \mathrm{d}\mu(\theta) + \epsilon \int_{\mathbb{R}^{K_d}} \nabla \bar{v}(\theta, t) \cdot \nabla w(\theta) \mathrm{d}\mu(\theta) = 0 \,, \text{ for all } w \in H^1(\mathrm{d}\mu) \,. \tag{10.31}$$

In particular taking $w = 1$ shows that the mean

$$\bar{\bar{v}} := \int_{\mathbb{R}^{K_d}} \hat{v}(\theta, t) \mathrm{d}\mu(\theta) = \int_{\mathbb{R}^{K_d}} G \mathrm{d}\mu \tag{10.32}$$

is constant in time.

We will use a Poincare inequality, i.e. that there is a positive constant γ such that

$$0 < \gamma := \inf_{\{w \in H^1(\mathrm{d}\mu) : \int w \mathrm{d}\mu = 0\}} \frac{\int_{\mathbb{R}^{K_d}} |\nabla w(\theta)|^2 \mathrm{d}\mu}{\int_{\mathbb{R}^{K_d}} w^2(\theta) \mathrm{d}\mu} \tag{10.33}$$

holds, see [3]. The energy estimate obtained by taking $w = \bar{v}$ in (10.31) and $c = \bar{\bar{v}}$ implies that $y(t) := \int_{\mathbb{R}^{K_d}} |\bar{v}(\theta, t)|^2 \mathrm{d}\mu/2$ satisfies

$$y'(t) + \epsilon\gamma y(t) \leq 0\,. \tag{10.34}$$

Therefore we have an exponential convergence rate towards the ergodic limit

$$\int_{\mathbb{R}^{K_d}} (\hat{v}(\theta, t) - \bar{\bar{v}})^2 \mathrm{d}\mu \leq e^{-\epsilon\gamma t} \int_{\mathbb{R}^{K_d}} (\hat{v}(\theta, 0) - \bar{\bar{v}})^2 \mathrm{d}\mu\,, \tag{10.35}$$

provided $\gamma > 0$ (which is implied by assumption (10.39)) and dominated convergence implies the pointwise convergence $\lim_{t\to\infty} \hat{v} - \bar{\bar{v}} = 0$ so that (10.32) yields

$$\lim_{t\to\infty} \mathbb{E}[g\big(y, \alpha(x, \theta_t)\big)] = \int_{\mathbb{R}^{K_d}} G \mathrm{d}\mu\,. \tag{10.36}$$

We are interested in the asymptotic limit of the Poincare constant γ as $\epsilon \to 0+$. Such Poincare inequalities have been proved by large deviation theory in [18] and potential theory in [3], which is closely related to the rare event result of Theorem 9.10. To describe these results, assume that G has isolated critical points and let the maximum elevation gain in the deepest valley for any continuous path $\omega_k : [0, 1] \to \mathbb{R}^{K_d}$ from a local minimum point θ_k to the global minimum point θ^* be denoted by

$$m := \sup_{\theta_k} \inf_{\{\omega_k \,:\, \omega_k(0) = \theta_k \,\&\, \omega_k(1) = \theta^*\}} \max_{s\in[0,1]} \big(G(\omega_k(s)) - G(\theta_k)\big) =: G(\theta_{12}) - G(\theta_1) \tag{10.37}$$

where $G(\theta_{12})$ is the highest saddle point and $G(\theta_1)$ the local minimum (i.e. not the global minimum) in the deepest valley, which is assumed to be unique. The work [18] shows that

$$-\lim_{\epsilon\to+0} \epsilon \log \gamma = m\,. \tag{10.38}$$

This result has been refined to include the G dependent prefactor C_G to the exponential $\gamma \simeq C_G e^{-m/\epsilon}$, see [3], provided that

$$\begin{aligned}
&G \in \mathcal{C}^3(\mathbb{R}^{K_d})\\
&\liminf_{|\theta|\to\infty} |\nabla G(\theta)| = \infty\,,\\
&\liminf_{|\theta|\to\infty} \big(|\nabla G(\theta)| - 2\Delta G(\theta)\big) = \infty\,,\\
&m \text{ and } \theta^* \text{ are unique}\,,\\
&\text{and that at the local minima and saddle points the Hessian of } G \text{ is non degenerate.}
\end{aligned} \tag{10.39}$$

Since $\alpha \in \mathcal{N}_K$ we have $G \in \mathcal{C}^\infty(\mathbb{R}^{K_d})$ and as the cost function for large $|\theta| > 2R_K$ is coercive also the other conditions in (10.39) hold. The small noise limit (10.37) shows that the lowest eigenvalue 0 of $\mathcal{L}$ is related to the global minimum point θ^* and the next eigenvalue γ is related to the minimum point θ_1 with the maximal elevation gain m. The rate γ implies that the number of stochastic gradient iterations n to approximate the minimum $\min_\theta G(\theta)$ with error δ becomes $n \sim (\epsilon\gamma\Delta s)^{-1} \log \delta^{-1}$, which would be exponentially large with respect to $\epsilon \sim \delta$, unless m would decrease as $K \to \infty$. In fact the neural network approximation has this property: the work [17] proves that $m = \mathcal{O}(K^{-1/d})$ for a half rectified network.

Let ρ be the initial probability density of θ_0. It is not the weighted $L^2(\mathrm{d}\mu)$ norm that is of primary interest but the error

$$\begin{aligned}
&\big|\mathbb{E}[g\big(y, \alpha(x, \theta_\tau)\big)] - \int_{\mathbb{R}^{K_d}} G \mathrm{d}\mu\big| = \big|\mathbb{E}[g\big(y, \alpha(x, \theta_\tau)\big)] - \lim_{\tau\to\infty} \mathbb{E}[g\big(y, \alpha(x, \theta_\tau)\big)]\big| \\
&= \big| \int_{\mathbb{R}^{K_d}} \bar{v}(\theta, \tau)\rho \mathrm{d}\theta\big| \leq \big(\int_{\mathbb{R}^{K_d}} \bar{v}^2(\theta, \tau) e^{-G/\epsilon} \mathrm{d}\theta\big)^{1/2} \big(\int_{\mathbb{R}^{K_d}} e^{G/\epsilon}\rho^2 \mathrm{d}\theta\big)^{1/2} \leq \mathcal{O}(e^{-\epsilon\gamma\tau/2 + c/\epsilon}) \,.
\end{aligned} \tag{10.40}$$

Here we have used (10.36), (10.35), Cauchy's inequality and the assumption that ρ has compact support. Convergence based on this estimate requires $\tau \gg e^{m/\epsilon}/\epsilon^2$, which also would grow exponentially with ϵ, unless m is small.

10.2.5 A priori bounds on $\mathbb{E}[|\bar{\theta}_n|^2]$ and $E[|\theta_t|^2]$

In the domain $|\theta| > \sqrt{2}R_K$, where R_K is the parameter in the penalty (10.4), we have $\nabla_\theta g\big(y, \alpha(x, \theta)\big) = 2(\theta - \theta\frac{y}{|\theta|})$ and $\nabla G(\theta) = 2\theta(1 - \frac{\mathbb{E}[y]}{|\theta|})$ which we use below to show

Lemma 10.4. *Assume that $|\bar{\theta}_0| \leq R_K$ and $|\theta_0| \leq R_K$, then there is a constant C, depending on R_K and K, such that for any $n \in \mathbb{N}$ and $t \in [0, \infty)$*

$$\begin{aligned}
\mathbb{E}[|\bar{\theta}_n|^2] &\leq \mathbb{E}[|\bar{\theta}_0|^2] + Cn\Delta s \,, \\
\mathbb{E}[|\theta_t|^2] &\leq \mathbb{E}[|\theta_0|^2] + Ct \,,
\end{aligned} \tag{10.41}$$

and a constant C_τ such that for $|\alpha| \leq 3$

$$\mathbb{E}\Big[\Big|\partial_{\bar{\theta}}^\alpha \mathbb{E}[g\big(y, \alpha(x, \theta_\tau)\big) \mid \theta_t = \bar{\theta}]\Big|\Big] \leq C_\tau \,.$$

Proof. We have for Δs sufficiently small

$$\begin{aligned}
&\mathbb{E}[|\bar{\theta}_{n+1}|^2] = \mathbb{E}[|\bar{\theta}_n - \nabla_\theta g\big(y, \alpha(x, \theta)\big)\Delta s + \sqrt{2\epsilon}\Delta W_n|^2] = \mathbb{E}[|\bar{\theta}_n|^2 \\
&\quad + \Delta s \mathbb{E}\Big[(\mathbf{1}_{|\theta_n| > \sqrt{2}R_K} + \mathbf{1}_{|\theta_n| \leq \sqrt{2}R_K})\big(-2\bar{\theta}_n \cdot \nabla g\big(y, \alpha(x, \bar{\theta}_n)\big) + |\nabla g\big(y, \alpha(x, \bar{\theta}_n)\big)|^2 \Delta s\big)\Big] \\
&\quad + \mathbb{E}[2\epsilon|\Delta W_n|^2] - 2\sqrt{2\epsilon}\mathbb{E}[\Delta W_n \cdot \nabla g\big(y, \alpha(x, \bar{\theta}_n)\big)\Delta s] + 2\sqrt{2\epsilon}\mathbb{E}[\bar{\theta}_n \cdot \Delta W_n] \\
&\leq \mathbb{E}[|\bar{\theta}_n|^2] + \Delta s C - 2(\mathbb{E}[|\theta_n|^2] - \mathbb{E}[|\theta_n|]\mathbb{E}[|y|])\Delta s + (\Delta s)^2 \mathbb{E}[|\theta_n|^2 - 2y|\theta_n| + y^2] \\
&\leq \mathbb{E}[|\bar{\theta}_n|^2] + \Delta s C
\end{aligned}$$

and its evolution implies that for any $j \in \mathbb{N}$ and $|\bar\theta_0| \le R_K$ and $\Delta s C \le 1$

$$\mathbb{E}[|\bar\theta_j|^2] \le 2R_K^2 + Cj\Delta s\,.$$

Similarly we have by Ito's formula for $|\theta_t| > \sqrt{2}R_K$ that

$$\mathrm{d}|\theta_t|^2 = -2\theta_t \cdot \nabla_\theta G(\theta_t)\mathrm{d}t + 2\sqrt{2\epsilon}\theta_t \cdot \mathrm{d}W_t + 2\epsilon K(d+2)\mathrm{d}t \tag{10.42}$$

so that if $|\theta_0| \le R_K$

$$\mathbb{E}[|\theta_t|^2] \le 2R_K^2 + Ct\,. \tag{10.43}$$

The combination of (10.42) and (10.43) shows that there is a constant C' such that

$$|\mathbb{E}\big[\mathbb{E}[g\big(y,\alpha(x,\theta_\tau)\big) \mid \theta_t = \bar\theta_t]\big]| \le C(1+\mathbb{E}\big[\mathbb{E}[|\theta_\tau|^2 \mid \theta_t = \bar\theta_t]\big]) \le C'(1+\epsilon\tau)\,.$$

To estimate derivatives of the function $v(\theta,t) = \mathbb{E}[g\big(y,\alpha(x,\theta_\tau)\big) \mid \theta_t = \theta]$ we use stochastic flows, i.e. derivatives of the process θ_t with respect to changes in θ_s for $s < t$. Let $\theta'_{ij}(t,s) := \frac{\partial\theta_i(t)}{\partial\theta_j(s)}$ be the first variation. Differentiation of (10.22) yields

$$\mathrm{d}\theta'(t,s) = -G''(\theta_t)\theta'(t,s)\mathrm{d}t$$

and similarly for the higher variations $\theta''(t,s) := \frac{\partial^2\theta_t}{\partial\theta_s^2}$

$$\begin{aligned}
\mathrm{d}\theta''_{\cdot\cdot k}(t,s) &= -G''(\theta_t)\theta''_{\cdot\cdot k}(t,s)\mathrm{d}t - \sum_j G'''_{\cdot\cdot j}\theta'_{\cdot\cdot}(t,s)\theta'_{jk}(t,s)\,,\\
\mathrm{d}\theta'''_{\cdot\cdot k\ell}(t,s) &= -G''(\theta_t)\theta'''_{\cdot\cdot k\ell}(t,s)\mathrm{d}t - \sum_j G'''_{\cdot\cdot j}\theta'_{\cdot\cdot}(t,s)\theta''_{jk\ell}(t,s)\\
&\quad - \sum_j G'''_{\cdot\cdot j}\theta''_{\cdot\cdot\ell}(t,s)\theta'_{jk}(t,s) - \sum_{jj'} G''''_{\cdot\cdot jj'}\theta'_{\cdot\cdot}(t,s)\theta'_{jk}(t,s)\theta'_{j'\ell}(t,s)\,.
\end{aligned} \tag{10.44}$$

The definition $z(t,s) = (\theta'(t,s), \theta''(t,s), \theta'''(t,s))$ makes it possible to write the dynamics (10.44) as $\mathrm{d}z_t = U(z_t,\theta_t)\mathrm{d}t$. We have

$$\partial_{\theta_k} v(\theta,t) = \sum_j \mathbb{E}[\partial_{\theta_j} g\big(y,\alpha(x,\theta_\tau)\big)\theta'_{jk}(\tau,t) \mid \theta_t = \theta]$$

with similar expressions based on z for the higher derivatives. Since $g(y,\alpha(x,\cdot))$ and U are Lipschitz continuous we obtain by (10.41) that $\mathbb{E}[|\nabla_\theta v(\bar\theta_t,t)|]$ is bounded by a constant depending on τ. Analogously also the expected value of the second and third derivatives of v are bounded. □

10.2.6 The $\mathcal{O}(\Delta s)$ estimate

Lemma 10.5. *Write (10.25) as $\mathrm{d}\bar\theta_s = \bar a(s,\bar\theta)\mathrm{d}s + \bar b\,\mathrm{d}W_s$ and assume that there is a constant C such that for $|\alpha| \le 2$ the functions $c : [0,\infty)\times\mathbb{R}^{K_d} \to \mathbb{R}$ and $u : [0,\infty)\times\mathbb{R}^{K_d} \to \mathbb{R}$ satisfy*

$$\begin{aligned}
\mathbb{E}[|\partial^\alpha c(s,\bar\theta_s)|^2 + |\partial^\alpha u(s,\bar\theta_s)|^2] + \min(\mathbb{E}[|\bar a u(s,\bar\theta_s)|^2, \mathbb{E}[|\bar a\nabla c(s,\bar\theta_s)|^2) \le C\,,\\
\min(\mathbb{E}[|\bar a(s,\bar\theta_s)(c(s,\bar\theta_s) - \bar c(s,\bar\theta_s))|^2], \mathbb{E}[|\bar a(s,\bar\theta_s)\nabla u(s,\bar\theta_s)|^2] \le C\,,
\end{aligned}$$

and let

$$
\begin{aligned}
&s_n := n\Delta s\,,\\
&s_n \le s < s_{n+1}\,,\\
&\text{for any function } c : [0,\infty)\times\mathbb{R}^K \to \mathbb{R} \text{ denote } \bar{c}(s,\bar{\theta}) := c(s_n,\bar{\theta}_n)\,,
\end{aligned}
$$

then there holds

$$
h(s) \;:=\; \mathbb{E}\Big[\Big(\bar{c}(s,\bar{\theta}) - c\big(s,\bar{\theta}(s)\big)\Big)u\big(s,\bar{\theta}(s)\big)\Big] = \mathcal{O}(\Delta s)\,.
$$

Proof. Since $\bar{c}(s,\bar{\theta}) = c(s_n,\bar{\theta}(s_n))$, we have

$$
h(s_n) = \mathbb{E}[\big(\bar{c}(s_n,\bar{\theta}) - c(s_n,\bar{\theta}(s_n))\big)u(s_n,\bar{\theta}(s_n))] = 0. \tag{10.45}
$$

Provided $|h'(s)| \le C$, the initial condition (10.45) implies that $h(s) = \mathcal{O}(\Delta s)$, for $s_n \le s < s_{n+1}$. Therefore, it remains to show that $|h'(s)| \le C$. Let $\beta(s,x) := -\Big(c(s,x) - c\big(s_n,\bar{\theta}(s_n)\big)\Big)u(s,x)$, so that $h(s) = \mathbb{E}[\beta(s,\bar{\theta}(s))]$. Then by Ito's formula

$$
\begin{aligned}
\frac{\mathrm{d}h}{\mathrm{d}s} = \frac{\mathrm{d}}{\mathrm{d}s}\mathbb{E}\left[\beta(s,\bar{\theta}(s))\right] &= \mathbb{E}\left[\mathrm{d}\beta(s,\bar{\theta}(s))\right]/\mathrm{d}s\\
&= \mathbb{E}\left[\left(\beta'_s + \bar{a}\nabla_\theta\beta + \mathrm{trace}(\frac{\bar{b}\bar{b}^T}{2}\beta'')\right)\mathrm{d}s + \nabla_\theta\beta\cdot\bar{b}\,\mathrm{d}W_s\right]/\mathrm{d}s\\
&= \mathbb{E}\left[\beta'_s + \bar{a}\nabla_\theta\beta + \mathrm{trace}(\frac{\bar{b}\bar{b}^T}{2}\beta'')\right]\\
&= \mathcal{O}(1).
\end{aligned}
$$

Therefore there exists a constant C such that $|h'(s)| \le C$, for $s_n < s < s_{n+1}$, and consequently

$$
h(s) = \mathbb{E}[\big(\bar{c}(s,\bar{\theta}) - c(s,\bar{\theta}(s))\big)u(s,\bar{\theta}_s)] = \mathcal{O}(\Delta s), \quad \text{for } s_n \le s < s_{n+1}.
$$

□

Chapter 11

Appendices

11.1 Tomography Exercise

Tomographic imaging is used in medicine to determine the shape/image of a bone or interior organ. One procedure for doing this is by projecting X-rays from many different angles through the body (see figure 1), measure the strength of the X-rays that has gone through the image, and compute how the image has to be to comply with the X-ray output. Reconstructing an image this way is called tomographic reconstruction, and it is the problem we look at in this project.

In our case we first superimpose a grid over the image we wish to perform tomographic imaging on to an $n \times n$ pixel image represented with image values as vector $(f_i)_{i=1}^{n^2}$. The image values are assumed to be constant within each cell of the grid. An $n = 3$ case with vertical and horizontal projections serves the purpose of further explaining the problem: In figure 2 we have superimposed a 3×3 square grid on the image $f(x, y)$. The rays are the lines running through the $x - y$ plane (we disregard the width of the lines here assuming they are all of the same width and very thin). The projections are given the representation p_i, we say that p_i is the ray sum measured with the ith ray. The relationship between the f_j's and the p_i's may be expressed as the set of linear equations

$$\sum_{j=1}^{n^2} A_{ij} f_j = p_i, \quad i = 1, \ldots, n. \tag{11.1}$$

For example, the first equation in the 3×3 case only goes through f_1, f_4 and f_7 yielding the equation

$$A_{11} f_1 + A_{14} f_4 + A_{17} f_7 = p_1,$$

The linear system of equations created by the horizontal and vertical projections in figure

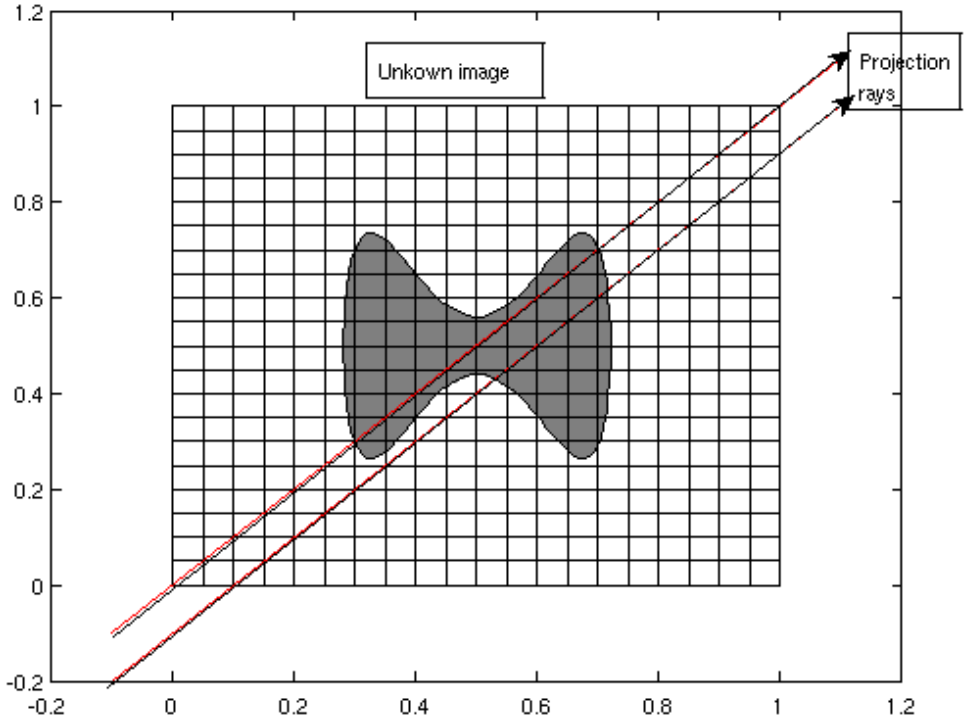


Figure 11.1: Illustration of tomographic imaging. The image on the unit square represents our unknown image which we send rays through to determine.

2 written on the form $An = p$ is

$$\begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \\ f_6 \\ f_7 \\ f_8 \\ f_9 \end{pmatrix} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \\ p_6 \end{pmatrix} \tag{11.2}$$

In this case, $A \in \mathbb{R}^{6\times 9}$. The problem is underdetermined so the least squares way of solving this problem:

$$f = (A^T A)^{-1} A^T p, \tag{11.3}$$

fails because $A^T A$ is singular. One way to deal with the singular matrix is to instead

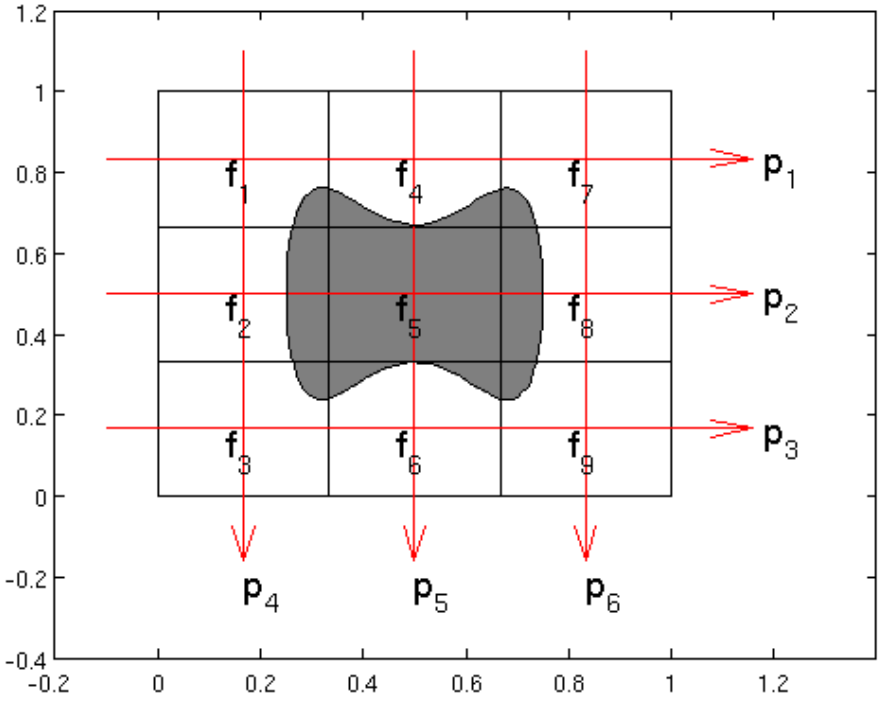


Figure 11.2: Illustration of horizontal and vertical projections on a 3×3 image.

solve

$$f = (A^T A + \delta I_{n^2})^{-1} A^T p,$$

where δ is a small number.

Exercise 1.

Download the image "ImageEx1.jpg" and the matlab program "rayItHorVert.m". This image is our unknown image (we only have the solution to compare). Create an image matrix by the command

```
image = imread('ImageEx1.tif')
```

Create a projection vector of the image by calling

```
p=rayItHorVert(f)
```

Write a matlab program that takes as input a vector $p \in \mathbb{R}^{6\times 1}$, creates the matrix $A \in \mathbb{R}^{6\times 9}$ given in (11.2) (for $n = 3$) and finds the tomographically reconstructed image f by the computation (11.3). Use

```
f=reshape(f,n,n)
```

to reshape the vector f into an $n \times n$ matrix and plot by the commands

```
colormap(gray)
imagesc(f)
```

Also plot the matrix "image" and compare results. As a reference, the result should look like figure 3:

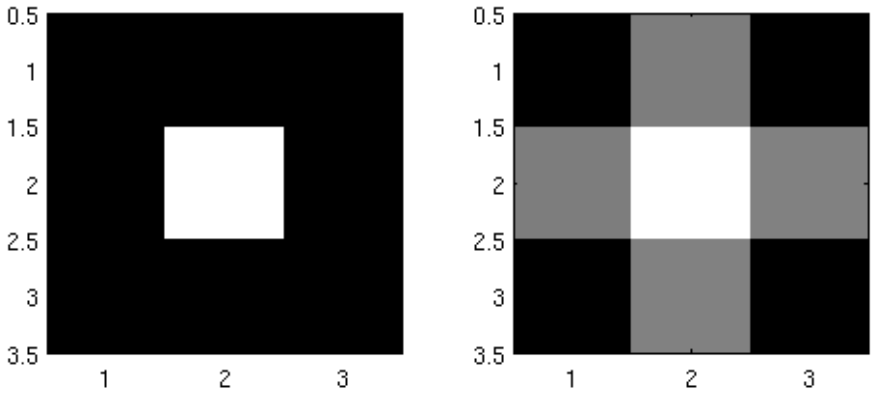


Figure 11.3: Illustration of the image "ImageEx1.jpg" (left) and the tomographic reconstruction (right).

Hint: The matrix A can be created quite easily with the Kronecker product $\otimes$ which is defined as follows:

$$B \otimes C = \begin{pmatrix} BC_{11} & BC_{12} & \dots & BC_{1n} \\ BC_{21} & BC_{22} & \dots & BC_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ BC_{m1} & BC_{m2} & \dots & BC_{mn} \end{pmatrix} \tag{11.4}$$

where $C \in \mathbb{R}^{m \times n}$ and B is an arbitrary matrix. In matlab the operation $B \otimes C$ is written

```
kron(B,C)
```

Exercise 2.

Use the hint in exercise 1. to generalize the matlab program to work for any n value. That is, write a program that takes as input an n-value and a vector $p \in \mathbb{R}^{2n \times 1}$, and creates a matrix $A \in \mathbb{R}^{2n \times n^2}$ with similar structure as the one in (11.2).

(a)

Download the image "Ball.tif" and solve the problem as in exercise one. One might improve the reconstructed image quality by filtering the image. Implement a scheme which removes values below a certain threshold in the matrix f and plot the result.

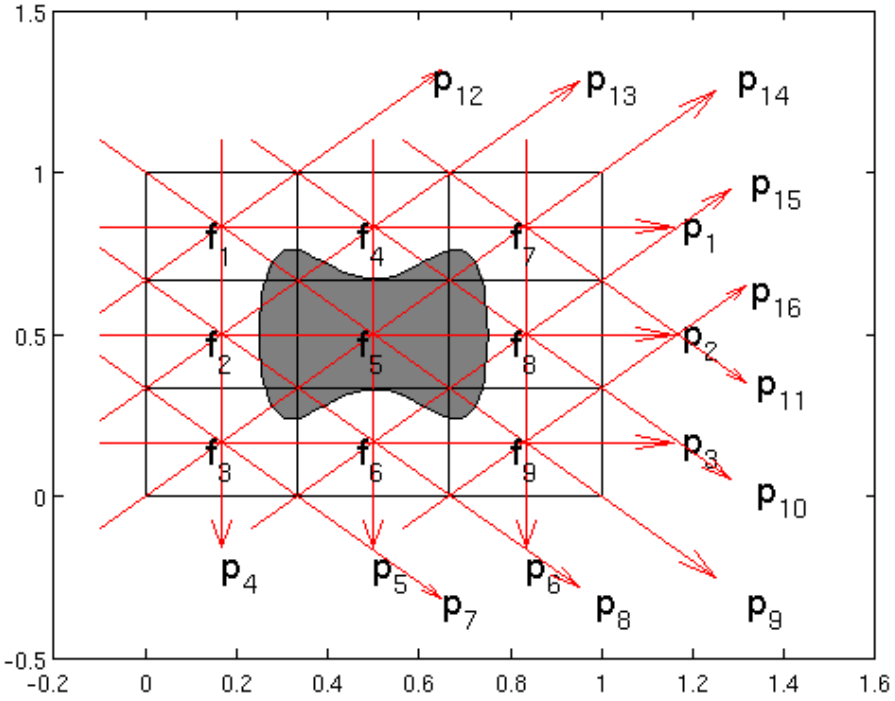


Figure 11.4: Illustration of horizontal, vertical and diagonal projections on a 3×3 image.

(b)

Assume that you have X-rayed a square shaped suitcase containing a circular shaped bomb. The file "pVector.mat" consists of the projection vector which you read by the command

```
load('pVector.mat')
```

What is approximately the position of the bomb? (Assume unit square coordinates).

(c)

Download the image "TwoBalls.tif" and solve the problem as in exercise one. Why does the reconstructed image differ so strongly from the real one?

The scheme implemented in exercise 3 improves the reconstructed image.

Exercise 3. - Week project exercise

The next step is to add more projections to our tomographic imaging. As illustrateted in figure 4, we use horizontal, vertical and diagonal projections. For the $n = 3$ case the

linear set of equations $Af = p$ is

$$\begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \\ f_6 \\ f_7 \\ f_8 \\ f_9 \end{pmatrix} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \\ p_6 \\ p_7 \\ p_8 \\ p_9 \\ p_{10} \\ p_{11} \\ p_{12} \\ p_{13} \\ p_{14} \\ p_{15} \\ p_{16} \end{pmatrix} \tag{11.5}$$

Write a program that takes as input an n-value and a vector $p \in \mathbb{R}^{(6n-2)\times 1}$, and creates a matrix $A \in \mathbb{R}^{(6n-2)\times n^2}$ with similar structure to the one in (11.5). Download the image "TwoBalls.tif" and the program "rayItHorVertDiag.m" which you use to create the projection vector by the command

```
p=rayItHorVertDiag(f)
```

Solve this image problem as in exercise 2 (c). Implement the filtering technique here as well. Compare this reconstruction to the one in 2 (c).

Exercise 4. - Week project exercise

The reason we are looking at low resolution images above is that for an $n \times n$ image the matrix $A \in \mathbb{R}^{(6n-2)\times n^2}$. This means that $A^T A \in \mathbb{R}^{n^2 \times n^2}$ which is so huge, that even for relatively small n that we can not solve the problem (11.3) in Matlab the way we have done in the exercises above. The paper "Algebraic reconstruction algorithms" describes an iterative algorithm solving the tomographic reconstruction problem which works for higher resolution images (see page 278). Read the first pages of this paper and implement this algorithm using horizontal, vertical and diagonal projections as in exercise 3. Try your algorithm on the picture "Pear.tif"

11.2 Molecular Dynamics

Here some discussion about the MD code will appear.

```
#include <math.h>
#include <stdio.h>
```

```
#include <stdlib.h>
#include <iostream>
#include <iomanip>
#include <fstream>
#include <sstream>
#include <string>

//
// Compile with g++ -O2 -o main main.cpp
//

using namespace std;

// ---------- Definitions ----------
typedef double real;
real sqr(real n){return n*n;}
enum BoundaryCond {periodic, flow};

// ---------- Cell and particle structures ----------
struct Parameters
{
  real sigma, epsilon, cutoff, dt, T, temp, size[3];
  int cells[3], cellsTot;
  BoundaryCond bc;
};

struct Particle
{
  real m;
  real x[3];
  real v[3];
  real F[3];
  real Fold[3];
  int flag;

};

struct ParticleList
{
  Particle p;
  ParticleList *next;
};

typedef ParticleList* Cell;
```

```
void insertListElement(ParticleList **root, ParticleList *pl)
{
  pl->next = *root;
  *root = pl;
}

void deleteListElement(ParticleList **pl)
{
  *pl = (*pl)->next;
}

int index(int *i, int *cells)
{
  return i[0] + cells[0]*(i[1] + cells[1]*i[2]);
}

// ---------- Function definitions ----------
void inputParameters(Parameters&);
void initData(Cell*, Parameters&);
void integrate(real, Cell*, Parameters&);
void compF(Cell*, Parameters&);
void compX(Cell*, Parameters&);
void compV(Cell*, Parameters&);
real compE(Cell*, Parameters&);
void updateX(Particle*, real);
void updateV(Particle*, real);
void forceLJ(Particle*, Particle*, real, real);
void sortParticles(Cell*, Parameters&);
void saveParticles(Cell*, real, Parameters&);
void boltzmann(Particle*, real);
real gaussDeviate();

// ---------- Program and functions ----------
int main(int argc, char **argv)
{

  int s = system("rm -rf ./data/*.txt");
  Parameters p;
  inputParameters(p);
  Cell *grid = new Cell[p.cellsTot];
  //for (int i=0; i<p.cellsTot; ++i)
  //  grid[i] = NULL;
  initData(grid, p);
```

```
  saveParticles(grid, 0, p);
  integrate(0, grid, p);
  return s;

}

void inputParameters(Parameters &p)
{
  // Lennard-Jones parameters
  p.sigma = 3.4;
  p.epsilon = 120;
  // Box size
  for (int d=0; d<3; ++d)
    p.size[d] = 150*p.sigma;
  // Cells
  p.cutoff = 2.5*p.sigma;
  for (int d=0; d<3; ++d)
    p.cells[d] = (int) floor(p.size[d] / p.cutoff);
  p.cellsTot = 1;
  for (int d=0; d<3; ++d)
    p.cellsTot *= p.cells[d];
  // Timescale
  p.T = 20;
  p.dt = 1e-2;
  // Boundary condition
  p.bc = flow;
  // Save to file
  FILE *file = fopen("./data/parameters.txt", "w");
  fprintf(file, "%f %f %f %f %f ", p.sigma, p.epsilon, p.cutoff, p.T, p.dt);
  for (int d=0; d<3; ++d)
    fprintf(file, "%f ", p.size[d]);
  for (int d=0; d<3; ++d)
    fprintf(file, "%d ", p.cells[d]);
  fclose(file);
}

void initData(Cell *grid, Parameters &p)
{

  // Box 1
  real mass = 39.95;
  int n1 = 10, n2 = 10, n3 = 10;
  grid[0] = NULL;
  ParticleList **root = &grid[0];
```

```
for (int i=0; i<=2*n1; ++i)
  for (int j=0; j<=2*n2; ++j)
    for (int k=0; k<=2*n3; ++k)
      {
        // Face centered cubic
        if ( !((i+j+k)%2) )
          {
            ParticleList *pl = new ParticleList;
            pl->p.m = mass;
            pl->p.x[0] = 0.5*p.size[0] + (i-n1)*pow(2, 1.0/6.0)*p.sigma;
            pl->p.x[1] = 0.5*p.size[1] + (j-n2)*pow(2, 1.0/6.0)*p.sigma;
            pl->p.x[2] = 0.6*p.size[2] + (k-n3)*pow(2, 1.0/6.0)*p.sigma;
            pl->p.v[0] = 0;
            pl->p.v[1] = 0;
            pl->p.v[2] = -20.4;
            pl->p.flag = 0;
            insertListElement(root, pl);
          }
      }

// Box 2
n1 = 30, n2 = 30, n3 = 10;
for (int i=0; i<=2*n1; ++i)
  for (int j=0; j<=2*n2; ++j)
    for (int k=0; k<=2*n3; ++k)
      {
        // Face centered cubic
        if ( !((i+j+k)%2) )
          {
            ParticleList *pl = new ParticleList;
            pl->p.m = mass;
            pl->p.x[0] = 0.5*p.size[0] + (i-n1)*pow(2, 1.0/6.0)*p.sigma;
            pl->p.x[1] = 0.5*p.size[1] + (j-n2)*pow(2, 1.0/6.0)*p.sigma;
            pl->p.x[2] = 0.4*p.size[2] + (k-n3)*pow(2, 1.0/6.0)*p.sigma;
            pl->p.v[0] = 0;
            pl->p.v[1] = 0;
            pl->p.v[2] = 0;
            pl->p.flag = 1;
            insertListElement(root, pl);
          }
      }

// Noise
for (ParticleList *pl=grid[0]; pl!=NULL; pl=pl->next)
```

```
    boltzmann(&pl->p, 1.0);
  sortParticles(grid, p);
}

void boltzmann(Particle *p, real factor)
{
  for (int d=0; d<3; ++d)
    p->v[d] += factor * gaussDeviate();

}
real gaussDeviate()
{
  real a1, a2, s, r, b1;
  static int iset = 0;
  static real b2;
  if (!iset)
    {
      do {
        a1 = 2.0 * rand() / (RAND_MAX + 1.0) - 1.0;
        a2 = 2.0 * rand() / (RAND_MAX + 1.0) - 1.0;
        r = a1 * a1 + a2 * a2;
      } while (r>=1.0);
      s = sqrt(-2.0 * log(r) / r);
      b1 = a1 * s;
      b2 = a2 * s;
      iset = 1;
      return b1;
    }
  else
    {
      iset = 0;
      return b2;
    }
}
void integrate(real t, Cell *grid, Parameters &p)
{
  compF(grid, p);
  while (t < p.T-1e-9)
    {
      t += p.dt;
      compX(grid, p);
      compF(grid, p);
      compV(grid, p);
      saveParticles(grid, t, p);
```

```
      cout << scientific <<
        "t = " << t << " E = " << compE(grid, p) << endl;
    }
}

void compF(Cell *grid, Parameters &p)
{
  int* cells = p.cells;
  int i[3], j[3];
  // Loop over cells in each dimension
  for (i[0]=0; i[0]<cells[0]; i[0]++)
    for (i[1]=0; i[1]<cells[1]; i[1]++)
      for (i[2]=0; i[2]<cells[2]; i[2]++)
        // Loop over particles in each cell
        for (ParticleList *pl1=grid[index(i,cells)]; pl1!=NULL; pl1=pl1->next)
          {
            for (int d=0; d<3; ++d)
              pl1->p.F[d] = 0;
            // Loop over neighbours in each dimension
            for (j[0]=i[0]-1; j[0]<=i[0]+1; j[0]++)
              for (j[1]=i[1]-1; j[1]<=i[1]+1; j[1]++)
                for (j[2]=i[2]-1; j[2]<=i[2]+1; j[2]++)
                  {
                    bool outside = false;
                    int tmp[3];
                    if (p.bc==periodic)
                      {
                        // Periodic boundary
                        for (int d=0; d<3; ++d)
                          tmp[d] = j[d];
                        for (int d=0; d<3; ++d)
                          if (j[d]<0)
                            j[d] = cells[d]-1;
                          else if (j[d]>=cells[d])
                            j[d] = 0;
                      }
                    else if (p.bc==flow)
                      {
                        // Flow boundary
                        for (int d=0; d<3; ++d)
                          if (j[d]<0 || j[d]>=cells[d])
                            outside = true;
                      }
                    if (!outside)
```

```
                    {
                      // Check distance from particle pl1 to neighbour cell j
                      real dist = 0;
                      for (int d=0; d<3; ++d)
                        dist +=
                          sqr( min( pl1->p.x[d] - j[d] * 1.0 / cells[d],
                                    pl1->p.x[d] - (j[d] + 1) * 1.0 / cells[d] ) );
                      // Loop over particles in each neighbour cell
                      //if (dist<=p.cutoff)
                      for (ParticleList *pl2=grid[index(j,cells)]; pl2!=NULL; pl2=pl2->
                        if (pl1!=pl2)
                          {
                            real r = 0;
                            for (int d=0; d<3; ++d)
                              r += sqr(pl2->p.x[d] - pl1->p.x[d]);
                            if (r<=sqr(p.cutoff))
                              forceLJ(&pl1->p, &pl2->p, p.sigma, p.epsilon);
                          }
                      }
                    if (p.bc==periodic)
                      {
                        // Copy back
                        for (int d=0; d<3; ++d)
                          j[d] = tmp[d];
                      }
                  }
          }
}

void forceLJ(Particle *i, Particle *j, real sigma, real epsilon)
{
  real r = 0.0;
  for (int d=0; d<3; ++d)
    r += sqr(j->x[d] - i->x[d]);
  real s = sqr(sigma) / r;
  s = sqr(s) * s;
  real f = 24 * epsilon * s / r * (1 - 2 * s);
  for (int d=0; d<3; ++d)
    i->F[d] += f * (j->x[d] - i->x[d]);
}

void compX(Cell *grid, Parameters &p)
{
  int i[3];
```

```
  // Loop over cells in each dimension
  for (i[0]=0; i[0]<p.cells[0]; i[0]++)
    for (i[1]=0; i[1]<p.cells[1]; i[1]++)
      for (i[2]=0; i[2]<p.cells[2]; i[2]++)
        // Loop over particles in each cell
        for (ParticleList *pl=grid[index(i,p.cells)]; pl!=NULL; pl=pl->next)
          updateX(&pl->p, p.dt);

  // Update cells according to new positions
  sortParticles(grid, p);
}

void updateX(Particle *p, real dt)
{
  real a = dt * 0.5 / p->m;
  for (int d=0; d<3; ++d)
    {
      p->x[d] += dt * (p->v[d] + a * p->F[d]);
      p->Fold[d] = p->F[d];
    }
}

void compV(Cell *grid, Parameters &p)
{
  int i[3];
  // Loop over cells in each dimension
  for (i[0]=0; i[0]<p.cells[0]; i[0]++)
    for (i[1]=0; i[1]<p.cells[1]; i[1]++)
      for (i[2]=0; i[2]<p.cells[2]; i[2]++)
        // Loop over particles in each cell
        for (ParticleList *pl=grid[index(i,p.cells)]; pl!=NULL; pl=pl->next)
          updateV(&pl->p, p.dt);
}

void updateV(Particle *p, real dt)
{
  real a = dt * 0.5 / p->m;
  for (int d=0; d<3; ++d)
    {
      p->v[d] += a * (p->F[d] + p->Fold[d]);
    }
}

void sortParticles(Cell *grid, Parameters &p)
```

```
{
  int i[3], j[3];
  // Loop over cells in each dimension
  for (i[0]=0; i[0]<p.cells[0]; i[0]++)
    for (i[1]=0; i[1]<p.cells[1]; i[1]++)
      for (i[2]=0; i[2]<p.cells[2]; i[2]++)
        {
          // Pointers to particle list in cell i
          ParticleList **pl1 = &grid[index(i,p.cells)];
          ParticleList *pl2 = *pl1;
          // Traverse list in cell i
          while (pl2!=NULL)
            {
              bool outside = false;
              // Cell that particle belongs to
              for (int d=0; d<3; ++d)
                {
                  j[d] = (int) floor(pl2->p.x[d] * p.cells[d] / p.size[d]);
                  if (p.bc==periodic)
                    {
                      // Periodic boundary
                      if (j[d]<0)
                        j[d] = p.cells[d] - j[d] % p.cells[d];
                      else if (j[d]>=p.cells[d])
                        j[d] = j[d] % p.cells[d];
                    }
                  else if (p.bc==flow)
                    {
                      // Outflow boundary
                      if (j[d]<0 || j[d]>=p.cells[d])
                        outside = true;
                    }
                }
              // If not same cell
              if ( (i[0]!=j[0]) || (i[1]!=j[1])
                   || (i[2]!=j[2]) )
                {
                  // Delete particle from list
                  deleteListElement(pl1);
                  // Add to list in cell j
                  if (!outside)
                    insertListElement(&grid[index(j,p.cells)], pl2);
                }
              else
```

```
              pl1 = &pl2->next;
            pl2 = *pl1;
          }
      }
}

real compE(Cell* grid, Parameters &p)
{
  real e = 0;
  int i[3];
  // Loop over cells in each dimension
  for (i[0]=0; i[0]<p.cells[0]; i[0]++)
    for (i[1]=0; i[1]<p.cells[1]; i[1]++)
      for (i[2]=0; i[2]<p.cells[2]; i[2]++)
        // Loop over particles in each cell
        for (ParticleList *pl=grid[index(i,p.cells)]; pl!=NULL; pl=pl->next)
          {
            real v = 0;
            for (int d=0; d<3; ++d)
              v += sqr(pl->p.v[d]);
            e += 0.5 * pl->p.m * v;
          }
  return e;
}

void saveParticles(Cell* grid, real t, Parameters &p)
{
  stringstream ss;
  ss.str(""); ss << fixed << setprecision(6) <<  t/p.T;
  string fname("./data/" + ss.str() + ".txt");
  FILE *file = fopen(fname.c_str(), "w");

  int i[3];
  // Loop over cells in each dimension
  for (i[0]=0; i[0]<p.cells[0]; i[0]++)
    for (i[1]=0; i[1]<p.cells[1]; i[1]++)
      for (i[2]=0; i[2]<p.cells[2]; i[2]++)
        {
          // Loop over particles in each cell
          for (ParticleList *pl=grid[index(i,p.cells)]; pl!=NULL; pl=pl->next)
            {
              for (int d=0; d<3; ++d)
                fprintf(file, "%f ", pl->p.x[d]);
              for (int d=0; d<3; ++d)
```

```
                fprintf(file, "%f ", pl->p.v[d]);
              fprintf(file, "%d \n", pl->p.flag);
            }
        }
  fclose(file);
}
```

Chapter 12

Recommended Reading

The following references have been useful for preparing these notes and are recommended for further studies.

Stochastic Differential Equations

- Online material: [15]
- Numerics for SDE: [25, 29]
- SDE: [30]
- Advanced SDE: [24]

Probability

[11]

Mathematical Finance

- Basic stochastics for finance: [2]
- Finance in practice: [21]
- Finance with numerics: [40]

Partial Differential Equations

- Advanced PDE: [16]
- Online introduction: [14]
- FEM: [23]
- Advanced FEM: [4]
- Introductory DE and PDE: [13] and [37]

Variance Reduction for Monte Carlo Methods

[5]

Molecular Dynamics

[26], [6], [19]

Machine Learning

[22]

Bibliography

[1] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

[2] Andrew Baxter, Martinand Rennie. *Financial calculus : an introduction to derivative pricing*. Cambridge Univ. Press, Cambridge, 1996.

[3] Anton Bovier, Véronique Gayrard, and Markus Klein. Metastability in reversible diffusion processes. II. Precise asymptotics for small eigenvalues. *J. Eur. Math. Soc. (JEMS)*, 7(1):69–99, 2005.

[4] Susanne C. Brenner and L. Ridgway Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer-Verlag, New York, 1994.

[5] Russel E. Caflisch. Monte Carlo and quasi-Monte Carlo methods. In *Acta numerica, 1998*, volume 7 of *Acta Numer.*, pages 1–49. Cambridge Univ. Press, Cambridge, 1998.

[6] Eric Cancès, Mireille Defranceschi, Werner Kutzelnigg, Claude Le Bris, and Yvon Maday. Computational quantum chemistry: a primer. In *Handbook of numerical analysis, Vol. X*, Handb. Numer. Anal., X, pages 3–270. North-Holland, Amsterdam, 2003.

[7] Carlsson, Jesper , Sandberg, Mattias , and Szepessy, Anders . Symplectic pontryagin approximations for optimal design. *ESAIM: Mathematical Modelling and Numerical Analysis*, PREPRINT, 2008.

[8] M. G. Crandall, L. C. Evans, and P.-L. Lions. Some properties of viscosity solutions of Hamilton-Jacobi equations. *Trans. Amer. Math. Soc.*, 282(2):487–502, 1984.

[9] M G Crandall and P-L Lions. Viscosity solutions of hamilton-jacobi equations. *Trans. Amer. Math. Soc*, 277(1):1–42, 1983.

[10] Bruno Dupire. Pricing with a smile. *Risk*, 7(1):18–20, 1994.

[11] Richard Durrett. *Probability: theory and examples*. Duxbury Press, Belmont, CA, second edition, 1996.

[12] Heinz W. Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1996.

[13] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson. *Computational differential equations*. Cambridge University Press, Cambridge, 1996.

[14] L. C. Evans. An introduction to mathematical optimal control theory. http://math.berkeley.edu/∼evans/.

[15] L. C. Evans. An introduction to stochastic differential equations. http://math.berkeley.edu/∼evans/.

[16] Lawrence C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 1998.

[17] C. Daniel Freeman and Joan Bruna. Topology and geometry of half-rectified network optimization. *arXiv*, abs/1611.01540, 2016.

[18] Mark I. Freidlin and Alexander D. Wentzell. *Random perturbations of dynamical systems*, volume 260 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, Heidelberg, third edition, 2012. Translated from the 1979 Russian original by Joseph Szücs.

[19] Berend Frenkel, Daanand Smit. *Understanding molecular simulation : from algorithms to applications*. Academic, San Diego, Calif., 2. ed. edition, 2002.

[20] Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric numerical integration*, volume 31 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2002. Structure-preserving algorithms for ordinary differential equations.

[21] John C. Hull. *Options, futures, and other derivatives*. Prentice Hall International, London, 3. ed. edition, 1997.

[22] Yoshua Bengio Ian Goodfellow and Aaron Courville. *Deep learning*. MIT Press (2016), http://www.deeplearningbook.org, 2016.

[23] Claes Johnson. *Numerical solution of partial differential equations by the finite element method*. Studentlitteratur, Lund, 1987.

[24] Ioannis Karatzas and Steven E. Shreve. *Brownian motion and stochastic calculus*, volume 113 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition, 1991.

[25] Peter E. Kloeden and Eckhard Platen. *Numerical solution of stochastic differential equations*, volume 23 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1992.

[26] Claude Le Bris. Computational chemistry from the perspective of numerical analysis. *Acta Numer.*, 14:363–444, 2005.

[27] Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms. *arXiv*, abs/1511.06251, 2015.

[28] Jerrold E. Marsden and Shane D. Ross. New methods in celestial mechanics and mission design. *Bull. Amer. Math. Soc. (N.S.)*, 43(1):43–73 (electronic), 2006.

[29] G. N. Milstein. *Numerical integration of stochastic differential equations*, volume 313 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1995. Translated and revised from the 1988 Russian original.

[30] Bernt Øksendal. *Stochastic differential equations*. Universitext. Springer-Verlag, Berlin, fifth edition, 1998. An introduction with applications.

[31] Pablo Pedregal. Optimization, relaxation and Young measures. *Bull. Amer. Math. Soc. (N.S.)*, 36(1):27–58, 1999.

[32] Olivier Pironneau. *Optimal shape design for elliptic systems*. Springer Series in Computational Physics. Springer-Verlag, New York, 1984.

[33] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko. *The mathematical theory of optimal processes*. Translated by D. E. Brown. A Pergamon Press Book. The Macmillan Co., New York, 1964.

[34] Levent Sagun, Léon Bottou, and Yann LeCun. Singularity of the hessian in deep learning. *arXiv*, abs/1611.07476, 2016.

[35] M. Sandberg and A. Szepessy. Convergence rates of symplectic Pontryagin approximations in optimal control theory. *M2AN*, 40(1), 2006.

[36] Mattias Sandberg. Convergence rates for an optimally controlled ginzburg-landau equation, 2008.

[37] Gilbert Strang. *Introduction to applied mathematics*. Wellesley-Cambridge Press, Wellesley, MA, 1986.

[38] Andreĭ I. Subbotin. *Generalized solutions of first-order PDEs*. Systems & Control: Foundations & Applications. Birkhäuser Boston Inc., Boston, MA, 1995. The dynamical optimization perspective, Translated from the Russian.

[39] Curtis R. Vogel. *Computational methods for inverse problems*, volume 23 of *Frontiers in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002. With a foreword by H. T. Banks.

[40] Sam Wilmott, Pauland Howison and Jeff Dewynne. *The mathematics of financial derivatives : a student introduction*. Cambridge Univ. Press, Cambridge, 1995.

[41] L. C. Young. *Lectures on the calculus of variations and optimal control theory.* Foreword by Wendell H. Fleming. W. B. Saunders Co., Philadelphia, 1969.