Complexity and Error Analysis of Numerical Methods for Wireless Channels, SDE, Random Variables, and Quantum Mechanics

HÅKON ANDREAS HOEL

Doctoral Thesis
Stockholm, Sweden, 2012
Akademisk avhandling som med tillstånd av Kungl Tekniska högskolan framlägges till offentlig granskning för avläggande av Doktorgradsexamen onsdagen den 30 maj 2012 kl 10.15 i sal F3, Lindstedtsvägen 26, Kungliga Tekniska högskolan, Stockholm.

© Håkon Andreas Hoel, 2012

Tryck: E-print, www.eprint.se
Abstract

This thesis consists of four papers considering different aspects of stochastic process modeling, error analysis, and minimization of computational cost.

In Paper I, we construct a Multipath Fading Channel (MFC) model for wireless channels with noise introduced through scatterers flipping on and off. By coarse graining the MFC model a Gaussian process channel model is developed. Complexity and accuracy comparisons of the models are conducted.

In Paper II, we generalize a multilevel Forward Euler Monte Carlo method introduced by Giles [16] for the approximation of expected values depending on solutions of Itô stochastic differential equations. Giles work [16] proposed and analyzed a Forward Euler Multilevel Monte Carlo (MLMC) method based on realizations on a hierarchy of uniform time discretizations and a coarse graining based control variates idea to reduce the computational cost required by a standard single level Forward Euler Monte Carlo method. This work is an extension of Giles’ MLMC method from uniform to adaptive time grids. It has the same improvement in computational cost and is applicable to a larger set of problems.

In paper III, we consider the problem to estimate the mean of a random variable by a sequential stopping rule Monte Carlo method. The performance of a typical second moment based sequential stopping rule is shown to be unreliable both by numerical examples and by analytical arguments. Based on analysis and approximation of error bounds we construct a higher moment based stopping rule which performs more reliably.

In paper IV, Born-Oppenheimer dynamics is shown to provide an accurate approximation of time-independent Schrödinger observables for a molecular system with an electron spectral gap, in the limit of large ratio of nuclei and electron masses, without assuming that the nuclei are localized to vanishing domains. The derivation, based on a Hamiltonian system interpretation of the Schrödinger equation and stability of the corresponding hitting time Hamilton-Jacobi equation for non ergodic dynamics, bypasses the usual separation of nuclei and electron wave functions, includes caustic states and gives a different perspective on the Born-Oppenheimer approximation, Schrödinger Hamiltonian systems and numerical simulation in molecular dynamics modeling at constant energy.
Preface

This thesis consists of an introduction and the four following papers:


The author contributed to all sections of the paper.


The author contributed to the analysis and the description of the algorithms.


The author contributed to all sections of the paper.


The author contributed to the numerical examples and to a small degree to the analysis.
Acknowledgments

I would like to thank my supervisor Anders Szepessy for providing interesting and
challenging projects. My collaborators Christian Bayer, Ashraful Kadir, Henrik
Nyberg, Petr Plecháč, Mattias Sandberg, Erik von Schwerin, and Raúl Tempone.
A special thanks goes to Raúl Tempone for being a great host during my visits at his
university and to Christian Bayer for much needed help on sampling problems. My
colleagues Jesper Karlsson, Jonas Kiessling, Love Lindholm, and Georgios Zouraris
for discussions during the weekly meetings of professor Szepessy’s research group.

I would also like to thank all my colleagues at the numerical analysis department
at KTH, especially my kind friends Murtazo Nazarov, Jelena Popovic, and Sara
Zahedi.

This work is funded by Center for Industrial and Applied Mathematics (CIAM)
and King Abdullah University of Science and Technology, Saudi Arabia.
Contents

I  Introductory Chapters  1

1  Introduction  3

2  Probability background  9
  2.1  Random Variables and Probability Measures  9
  2.2  Convergence of Random Variables and Limit Theorems for Sampling  14

3  Monte Carlo Methods  17
  3.1  Examples and Properties  17
  3.2  Variance Reduction  21
  3.3  Sequential Stopping Rules  21

4  Adaptive Weak Approximations for SDE  25
  4.1  Stochastic Differential Equations (SDE)  25
  4.2  The Euler Method for Itô SDE  29
  4.3  Adaptive Weak Solution Approximation  35

5  Wireless Channel Modeling  39
  5.1  The Multipath Fading Channel  39
  5.2  From MFC Models to Gaussian Processes  45

6  Classical and Quantum Mechanics  47
  6.1  Classical Mechanics  47
  6.2  Quantum Mechanics  53

Bibliography  61
Part I

Introductory Chapters
Chapter 1

Introduction

If a system does not always produce the same output from a given initial state, we call it a non-deterministic system. Considering a non-deterministic system whose uncertainty is described in form of a density \( P \), mean value quantities on the form

\[
\int g(x)dP(x)
\]  
(1.0.1)

are often sought. For example:

- Let \( X \) be a random variable modeling the number of active users in a network with the with \( P(n) \) denoting the probability for having \( n \in \mathbb{N} \) active users (within a fixed time interval). Then (1.0.1) with \( g(n) = n \) represents the mean number of active users in the network.

- A mechanical system with \( N \) particles positioned at \( q \in \mathbb{R}^{3N} \), with momentum \( p \in \mathbb{R}^{3N} \), potential function \( V(q) \), and \( P(q,p) \) denoting the probability for the particles having the phase space configuration \( (q,p) \). Then the average potential energy of the system \( g(q) = V(q) \) is a quantity of interest.

In some settings, e.g., if the density \( P \) is not explicitly known or if it takes a subtle form, integrals on the form (1.0.1) have to be approximated. A typical way of approximating (1.0.1) is by Monte Carlo sampling

\[
\sum_{i=1}^{M} \frac{g(X_i)}{M},
\]

where the realizations \( X_i \) are generated according to or approximately according to the density \( P \). In this thesis, we will for four different non-deterministic systems study questions on how to efficiently generate realizations that are approximately distributed according to a density \( P \) and give estimates on weak approximation errors of type

\[
\left| \int g(x)dP(x) - \sum_{i=1}^{M} \frac{g(X_i)}{M} \right|
\]  
(1.0.2)
in terms of the computational cost. Let us be more specific.

Paper I

Wireless channel models model the received version of a signal transmitted wirelessly from a transmitter to a receiver. For industries developing wireless transmission equipment, such models are used to estimate the performance of equipment and software by simulations instead of more costly real world tests. But for such models to be of interest, they must be accurate and computationally efficient with respect to running time.

One of the most popular models of today, the Multipath Fading Channel model (MFC), is based on approximating the signal from superpositioning a finite number of contributing wave paths yielding an output signal on the form

\[ Y_t = \frac{1}{\sqrt{M}} \sum_{k=1}^{M} a_k e^{i \theta_k(t)}. \]

The superpositioning of wave paths is computationally costly, and, consequently, it is also costly to generate output signal realizations using MFC models. In Paper I, we propose a new MFC model with scatterers flipping on and off adding noise to the total signal, and by coarse graining this MFC model we derive a Gaussian process wireless channel model. Computational cost estimates in Paper I indicate that signal realizations are generated more efficiently by using a Gaussian process algorithm than by using an MFC algorithm.

Paper II

Stochastic Differential Equations (SDE) are non-deterministic processes whose future evolution is described by a probability distribution, as opposed to the deterministic evolution of an ordinary differential equation. Complex phenomena which might seem non-deterministic, such as stock market evolution, are frequently modeled by stochastic processes, cf. [5, 24].

In the second paper, we develop an adaptive Forward Euler Monte Carlo algorithm which for any sufficiently well behaved function \( g : \mathbb{R}^d \to \mathbb{R} \) approximates the expected value

\[ \mathbb{E}[g(X_T)] = \int g(X_T(\omega))dP(\omega), \tag{1.0.3} \]

where \( X_T \) is the solution of an Itô SDE. The algorithm we have developed constructs numerical realizations of the SDE on adaptive time grids using the adaptive Forward Euler method. To obtain variance reduction, realizations are constructed on different tolerance levels and the expected value (1.0.3) is approximated from the numerical realizations by the Monte Carlo method.

The multilevel Monte Carlo method, first introduced by Giles in [16], showed the improvement of computational cost of approximating \( \mathbb{E}[g(X_T)] \) with accuracy
$O(TOL)$ from $O(TOL^{-3})$ by a single level method to $O(TOL^{-2}\log(TOL)^2)$ by the multilevel method when the realizations $X_T(\omega)$ of the given SDE problem were generated numerically on uniform time grids. In Paper II we extend Giles’ multilevel method to the setting of adaptive time grid SDE realizations relying on the adaptive weak approximation methods for SDE developed by Szepessy et al. [30, 25, 26]. Our extended multilevel method is applicable to a larger set of SDE problems and, when comparable, it is shown to have the same computational cost as Giles’ method has.

**Paper III**

Given i.i.d. random variables $X_1, X_2, \ldots$ the typical way of approximating their expected value $\mu = E[X]$ using $M$ samples is the sample average

$$\overline{X}_M := \sum_{i=1}^{M} X_i/M.$$  

In Paper III, we consider the objective of choosing $M$ sufficiently large so that

$$P \left( |\overline{X}_M - \mu| > TOL \right) \leq \delta,$$  

(1.0.4)

for small, fixed accuracy-confidence constants $TOL > 0$ and $\delta > 0$. Clearly, $P(|\overline{X}_M - \mu| > TOL)$ decreases as $M$ increases, but at the same time the cost of computing $\overline{X}_M$ increases. From an application and cost point of view it is therefore of interest to have theory giving sharp bounds on the number of samples $M$ needed to fulfill (1.0.4). For some settings this exists. For example, if $E[|X|^\infty] < C$, it is possible to derive good theoretical upper bounds for $M$, but in the general case when no or little information of the distribution is given, however, little theory is known and the typical way of estimating $E[X]$ is by a sequential stopping rule; sequentially increasing the number of samples $M$ until the sampled moments fulfill a stopping criterion. In Paper III, we show that the “intuitive” stopping criterion

$$2 \left( 1 - \Phi \left( \frac{\sqrt{M}TOL}{\sigma_M} \right) \right) < \delta,$$

where

$$\Phi(z) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2} dx$$

gives a stopping rule that performs unreliably when sampling heavy-tailed r.v. From approximations of error bounds we construct a new stopping criterion based on second, third, and fourth order sample moments which according to numerical experiments performs more reliably and is only slightly more costly than the stopping rule with the “intuitive” stopping criterion.
CHAPTER 1. INTRODUCTION

Paper IV

Molecular dynamics is a computational method to study molecular systems in materials science, chemistry and molecular biology. The simulations are used, for example, in designing and understanding new materials or for determining biochemical reactions in drug design. The wide popularity of molecular dynamics simulations relies on the fact that in many cases it agrees very well with experiments. Indeed, given experimental data it is easy to verify correctness of the method by comparing with experiments at certain parameter regimes. However, if we want the simulation to predict something that has no comparing experiment, we need a mathematical estimate of the accuracy of the computation. In the case of molecular systems with few particles such studies are made by directly solving the Schrödinger equation. A fundamental and still open question in classical molecular dynamics simulations is how to verify the accuracy computationally, i.e., when the solution of the Schrödinger equation is not a computational alternative.

The aim of this paper is to derive qualitative error estimates for molecular dynamics and present new mathematical methods which could be used also for a more demanding quantitative accuracy estimation, without solving the Schrödinger equation. That is, let $\Phi$ be a solution of the time-independent Schrödinger equation

$$\left(-\frac{1}{2}M^{-1}\sum_{n=1}^{N}\Delta x_n + V\right)\Phi = E\Phi,$$

where $V$ is a given potential operator, $E \in \mathbb{R}$ is energy and $M$ is a the mass constant, and let $X(t)$ be a molecular dynamics path with total energy also equal to $E$. Then, our object of study is the approximation error of

$$\int_{\mathbb{R}^{3(N+n)}} g(X)\Phi(X,x)^*\Phi(X,x)\,dX\,dx \quad \text{and} \quad \lim_{T \to \infty} \frac{1}{T} \int_0^T g(X(t))\,dt$$

in terms of $M$ as $M \to \infty$, for any observable $g(X)$.

Having molecular dynamics error estimates opens, for instance, the possibility of systematically evaluating which density functionals or empirical force fields are good approximations and under what conditions the approximation properties hold. Computations with such error estimates could also give improved understanding when quantum effects are important and when they are not, in particular in cases when the Schrödinger equation is too computationally complex to solve.

Outline of the Introductory Chapters

In Chapter 2, we give a short description of random variables, probability distributions, and limit theorems in probability theory. In Chapter 3, we present topics on Monte Carlo methods such as variance reduction and dimension independent convergence rate. In Chapter 4, we give a short introduction to numerical methods and weak solution approximation for SDE. Chapter 5 presents statistical wave path
modeling of wireless channels, and in Chapter 6, we give a short outline on Classical and Quantum Mechanics.
Chapter 2

Probability background

Although the Events of Games, which Fortune solely governs, are uncertain, yet it may be certainly determin’d, how much one is more ready to lose than gain . . .

It is impossible for a Die, with such determin’d force and direction, not to fall on such determin’d side, only I don’t know the force and direction which makes it fall on such determin’d side, and therefore I call it Chance, which is nothing but the want of art.

—John Arbuthnot, Of the Laws of Chance 1.

2.1 Random Variables and Probability Measures

During renaissance times, loose guidelines involving reason, intuition, and observations were applied when assessing the uncertainty of evidence material in court, for discussing betting strategies in terms of odds, and for setting maritime insurance premiums. A first mathematical treatment of probability can be traced back to letters between Pierre de Fermat and Blaise Pascal in the year 1654 where they discuss, among other things, the division of stakes in fair gambling games. Let us therefore, in the spirit of gambling, develop our first random variable (r.v.) as a model of the game Heads or Tails. Heads or Tails is the game of predicting which side will face up when a coin is flipped. The game is quite successfully modeled by

\[
\text{Outcome} = \begin{cases} 
\text{Heads, with probability } p \\
\text{Tails, with probability } (1 - p),
\end{cases} \tag{2.1.1}
\]

1English translation (with additions) of De Ratiociniis in Ludo Aleae by Frans van Schooten and Christiaan Huygens (foreword, at least) in 1656.
where \( p \in [0, 1] \) models the coin’s proclivity towards Heads. (Typically with \( p = 1/2 \) when the coin is of fair shaped, or alternatively configured from experiments or measurements. For example, tossing the coin \( N \) times and setting \( p = \#\text{Heads}/N \).)

It is often possible to motivate both deterministic and stochastic models for the problem one is studying. In the case of Heads or Tails, for example, one might argue that if all input parameters—initial orientation, velocity and spin for the coin; air density; material and geometrical data for the coin and the ground a.s.f.—is known prior to the coin flip, the outcome of the coin flip ought to be deterministically predictable. But if the input data is uncertain and/or the deterministic model for the outcome is too complicated, a stochastic model might be preferable in comparison to a deterministic one. Furthermore, the choice of model should depend on the problem you are facing. For example, if you wish to estimate how likely ten consecutive Heads throws are, the stochastic model straightforwardly gives you the estimate \( p^{10} \), while if you wish to estimate a more specific property of the coin throws, for example the precise path of each coin flip, a deterministic model of the coin flips might be needed.

To present and analyze more complicated r.v. and stochastic models we introduce some terminology for probability spaces and r.v.

**Definition 2.1.1 (Probability space, \( \sigma \)-algebra, and Probability Measure)**

A probability space is a measure space triple \((\Omega, \mathcal{F}, P)\) with the sample space \( \Omega \), the event space \( \mathcal{F} \), and the probability measure \( P : \mathcal{F} \to [0, 1] \). The sample space is the set of outcomes, with an outcome signifying the result of single execution of the model. An event denotes the union of one or more outcomes, and the event space is the set of events. The event space \( \mathcal{F} \) is a \( \sigma \)-algebra, i.e., it is a collection of subsets of \( \Omega \) fulfilling the following:

(i) \( \mathcal{F} \) is non-empty.

(ii) Closed under complement: If \( A \in \mathcal{F} \), then \( A^C \in \mathcal{F} \), with \( A^C \) denoting the complement.

(iii) Closed under countable unions: If \( A_i \in \mathcal{F} \) for \( A_0, A_1, \ldots \), then \( \cup_i A_i \in \mathcal{F} \).

The probability measure acts on the measurable space \((\Omega, \mathcal{F})\) fulfilling the following criteria:

(i) Non-negative: \( P(A) \geq 0 \).

(ii) Countable additivity: For all collections a collection \( \{A_i\}_i \) contained in \( \mathcal{F} \) and pairwise disjoint,

\[
P(\cup_i A_i) = \sum_i P(A_i).
\]

(iii) Probability measure: \( P(\Omega) = 1 \).

A r.v. is a mapping defined on a probability space.
Definition 2.1.2 (Random Variable) Given the probability space \((\Omega, \mathcal{F}, P)\) and an arbitrary measurable space \((\Gamma, \mathcal{G})\), then \(X : \Omega \to \Gamma\) is said to be a random variable/vector if the map from \((\Omega, \mathcal{F})\) to \((\Gamma, \mathcal{G})\) is measurable, i.e.,

\[
\{ \omega : X(\omega) \leq G \} \in \mathcal{F}, \quad \forall G \in \mathcal{G}.
\]

A frequently encountered type of r.v. is the real-valued r.v. mapping from \((\Omega, \mathcal{F})\) to \((\mathbb{R}, \mathcal{B})\) where \(\mathcal{B}\) denoting the Borel \(\sigma\)-algebra. A real-valued r.v. \(X\) has a non-decreasing, right continuous Cumulative Distribution Function (CDF) \(F(x) := P(X \leq x)\) defined for \(x \in \mathbb{R}\) and a Probability Density Function (PDF) \(f(x) = P(X = x)\) which is connected to CDF by

\[
F(x) = \int_{-\infty}^{x} dF(s) = \int_{-\infty}^{x} f(s) \, ds.
\]

Often, r.v. are indirectly defined by their PDF, and we further note that moments such as the expected value

\[
\mu = E[X] := \int_{\mathbb{R}} x f(x) \, dx
\]

and the variance

\[
\sigma^2 = \text{Var}(X) := E[|X - E[X]|^2]
\]

are values often used both to describe and to analyze r.v.

Random Variables—Examples

We now include examples of some of the r.v. which appear in the papers of this thesis.

Example 2.1.3 (Heads or Tails a.k.a. Bernoulli distribution) Based on the model (2.1.1) let us define the real-valued r.v. \(X(\omega)\) as follows:

\[
X(\omega) = \begin{cases} 
1, & \text{if } \omega = \{\text{Heads}\} \\
0, & \text{if } \omega = \{\text{Tails}\}
\end{cases}
\] (2.1.2)

It has the probability space

\[
\Omega = \{\text{Heads}, \text{Tails}\}, \quad \mathcal{F} = \{\emptyset, \{\text{Heads}\}, \{\text{Tails}\}, \Omega\},
\]

and \(P(X = 1) = p = P(X = \{\text{Tails}\})\),

the PDF

\[
f(x) = \frac{\delta_x + \delta_{1-x}}{2},
\]

with \(\delta_x\) denoting the Dirac delta distribution, \(\mu = p\), and \(\sigma^2 = p(1 - p)\). (Strictly speaking, the function \(f\) is called a probability mass function when it only attains non-zero values at a countable number of \(x \in \mathbb{R}\), not a PDF.)
Example 2.1.4 (The Poisson distribution)  The Poisson distribution models the likelihood for a given number of identical independent events occurring within a fixed time interval. For example, the number of telephone calls in a call system or the number of wave scatterers becoming active in a wireless multi path fading channel, cf. Paper I. The PDF is given by

\[ f(k) = \frac{\lambda^k e^{-\lambda}}{k!} \] for \( k \in \mathbb{N}_0 \),

where \( \lambda > 0 \) is a positive parameter equaling the expected number of events on a given interval. That is, \( E[X] = \lambda \) and (it turns out that) also \( \text{Var}(X) = \lambda \).

Example 2.1.5 (Normal distribution)  The normal distribution is a very important distribution with a tremendous amount of applications. It is innately connected to the limit distribution of the scaled mean of samples of independent, identically distributed (i.i.d.) r.v. (see Section 2.2), and many real life uncertainty estimates depend on this distribution. Considering a normal r.v. \( X \) with mean \( \mu \in \mathbb{R} \) and variance \( \sigma^2 \in \mathbb{R}_+ \), the PDF is given by

\[ f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \]

We write \( X \sim \mathcal{N}(\mu, \sigma^2) \). We further note that the CDF of a standard normal r.v. \( X \sim \mathcal{N}(0,1) \), which often is used in estimates of sampling error, is in this thesis represented by \( \Phi(x) \).

Generalizations of the normal r.v. to \( \mathbb{R}^n \) and \( \mathbb{C}^n \) exist. In \( \mathbb{R}^n \), the multivariate normal with input parameters \( \mu \in \mathbb{R}^n \) and symmetric, positive definite covariance matrix \( K \in \mathbb{R}^{n \times n} \) has the PDF

\[ f(x) = \frac{1}{(2\pi)^{n/2} \det(K)^{1/2}} \exp \left( \frac{1}{2} (x - \mu)^T K^{-1} (x - \mu) \right). \]
Example 2.1.6 (Pareto distribution) The Pareto-distribution has the PDF

\[
f(x) = \begin{cases} \alpha x_m^\alpha x^{-(\alpha+1)} & \text{if } x \geq x_m \\ 0 & \text{else}, \end{cases} \tag{2.1.3}
\]

where \(\alpha, x_m \in \mathbb{R}_+\) are respectively the shape and the scale parameter. The moments of \(E[X^n]\) for the Pareto r.v. only exists for \(n < \alpha\) and, supposing \(\alpha > 2\), its mean and variance are given by

\[
\mu = \frac{\alpha x_m^\alpha}{\alpha - 1} \quad \text{and} \quad \sigma^2 = \frac{x_m^{2\alpha}}{(\alpha - 1)^2(\alpha - 2)}.
\]

The Pareto-distribution is a so called heavy-tailed distribution, meaning that its tail is not exponentially bounded, i.e.

\[
\lim_{x \to \infty} e^{\lambda x} P(X \geq x) = \infty, \quad \forall \lambda \in \mathbb{R}_+.
\]
CHAPTER 2. PROBABILITY BACKGROUND

The Italian economist Vilfredo Pareto introduced the Pareto distribution as model for income distribution. The pareto index \( \alpha \) models the breadth of income; the larger the pareto index, the smaller the proportion of high-income people.

2.2 Convergence of Random Variables and Limit Theorems for Sampling

In this section we review relevant convergence definitions and results from probability theory. For simplicity, the results are presented for sequences of scalar, real-valued r.v.

Definition 2.2.1 (Convergence in distribution) A sequence of real-valued r.v. \( X_0, X_1, \ldots \) converges in distribution to the r.v. \( X \) with the CDF \( F(x) \) if

\[
\lim_{n \to \infty} P(X_n \leq x) = F(x)
\]
at all continuity points \( x \in \mathbb{R} \) of \( F \).

Definition 2.2.2 (Convergence in probability) A sequence of real-valued r.v. \( X_0, X_1, \ldots \) converges in probability to the r.v. \( X \) if for any \( \epsilon > 0 \),

\[
\lim_{n \to \infty} P(|X_n - X| > \epsilon) = 0.
\]

We write \( X_n \to X \) in probability.

Definition 2.2.3 (Almost sure convergence) A sequence of real-valued r.v. \( X_0, X_1, \ldots \) converges almost surely to the r.v. \( X \) if

\[
P(\lim_{n \to \infty} X_n - X = 0) = 1.
\]

We write \( \lim_{n \to \infty} X_n = X \) a.s.

The strength relations between the convergence notions are

Almost sure conv. \( \implies \) Conv. in probability \( \implies \) Conv. in distribution.

Given a sequence of independent identically distributed (i.i.d.) r.v. \( \{X_i\}_{i=1}^n \), a question of interest is if and how fast the average \( S_n/n \), where \( S_n := \sum_{i=1}^M X_i \) converges to \( \mu = E[X_1] \) as the number of samples \( M \) increases. Before reviewing results regarding this question, we first recall the definition of independent r.v.

Definition 2.2.4 (Independent random variables) Two r.v. \( X_1 \) and \( X_2 \) mapping from \( (\Omega, \mathcal{F}) \) to \( (\mathbb{R}, \mathcal{B}) \) are independent if for all \( A_1, A_2 \in \mathcal{B} \)

\[
P(X_1 \in A_1, X_2 \in A_2) = P(X_1 \in A_1)P(X_2 \in A_2).
\]
2.2. CONVERGENCE OF RANDOM VARIABLES AND LIMIT THEOREMS FOR SAMPLING

In particular, independence implies that \( \mathbb{E}[X_1X_2] = \mathbb{E}[X_1] \mathbb{E}[X_2] \). The simplest asymptotic convergence results for \( S_n/n \) is the weak law of large numbers.

**Theorem 2.2.5 (Weak law of large numbers)** Suppose \( X_1, X_2, \ldots \) are i.i.d. r.v. with \( \mathbb{E}[|X_1|] < \infty \). Then \( S_n/n \to \mu \) in probability as \( M \to \infty \).

The result in the general form stated above follows from the Dominated Convergence Theorem, but let us prove it in the setting when \( \text{Var}(X_1) < \infty \) using Chebycheff’s inequality and independence: For any \( \epsilon > 0 \),

\[
P(|S_n/n - \mu| > \epsilon) \leq \frac{\mathbb{E}\left[ |S_n/n - \mu|^2 \right]}{\epsilon^2}
\]

\[
\leq \sum_{i=1}^{M} \frac{\text{Var}(X_i)}{M^2 \epsilon^2} \to 0 \text{ as } M \to \infty.
\]

The Strong law of large numbers is a, not surprisingly, a slightly stronger result telling us that in the limit \( n \to \infty \), the sample average \( S_n/n \) equals \( \mu \) on a measure 1 set (i.e., a.s.).

**Theorem 2.2.6 (Strong law of large numbers)** Suppose \( X_1, X_2, \ldots \) are i.i.d. r.v. with \( \mathbb{E}[|X_1|] < \infty \). Then \( S_n/n \to \mu \) a.s. as \( n \to \infty \).

The weak and strong laws of large numbers describes the asymptotic pointwise convergence of the sample mean \( S_n/n \). The Central Limit Theorem (CLT) gives you additional information on the asymptotic distribution of \( S_n/\sqrt{n} \).

**Theorem 2.2.7 (The Central Limit Theorem)** Suppose \( X_1, X_2, \ldots \) are i.i.d. r.v. with (as before) \( \mu = \mathbb{E}[X_1] \) and \( \sigma^2 = \text{Var}(X_1) < \infty \). Then

\[
\lim_{n \to \infty} P\left(\frac{S_n - \mu}{\sigma \sqrt{n}} \leq x\right) = \Phi(x),
\]

where \( \Phi(x) \) is the CDF of a standard normal distributed r.v. cf. Example 2.1.5.

An explicit \( n^{-1/2} \) convergence rate for the distributional convergence in the CLT was derived independently by Berry and Esseen in the 1940s through bounding multi-convoluted characteristic functions, cf. [4, 14].

**Theorem 2.2.8 (Berry-Esseen)** Suppose \( X_1, X_2, \ldots \) are i.i.d. r.v. with (as before) \( \mu = \mathbb{E}[X_1] \) and \( \sigma^2 = \text{Var}(X_1) \) and \( \rho = \mathbb{E}[|X_1 - \mu|^3] < \infty \). Then

\[
\sup_{x \in \mathbb{R}} |F_n(x) - \Phi(x)| \leq C \frac{\rho}{\sqrt{n} \sigma^3}
\]

where

\[
F_n(x) := P\left(\frac{S_n}{\sigma \sqrt{n}} \leq x\right).
\]

For multi-dimensional and other extensions of the above presented convergence results, see [12, 3].
Chapter 3

Monte Carlo Methods

The infinite we shall do right away. The finite may take a little longer.

—Stan Ulam

Monte Carlo methods are a class of algorithms relying on repeated sampling of r.v. to compute the quantities of interest. A simple example would be the sample average

$$g_M := \sum_{i=1}^{M} \frac{g(X_i)}{M},$$

(3.0.1)

for some sequence of i.i.d. r.v. $X_1, X_2, \ldots$ in $\mathbb{R}^n$, $g : \mathbb{R}^n \to \mathbb{R}$ and the number of samples $M$ predetermined or increased until the variance of $g_M$ is sufficiently small. MC methods are generally easy to implement, even for seemingly complicated problems, and their convergence rate is $\mathcal{O}(M^{-1/2})$, independent of the dimension of the problem. MC methods have become popular among statisticians and scientists in applied fields which apply the methods in estimations such as medicine efficiency, election predictions, and finance.

The field of MC methods grows by the day, so a clear cut definition of the term and summary of the subject becomes increasingly difficult. Here, we restrict ourselves to an informal presentation of the three central themes for MC methods through examples: how many samples $M$ are needed to estimate the quantity of interest reliably, what variance reduction is, and the dimension independent convergence rate of MC methods.

3.1 Examples and Properties

When approximating a quantity of interest by the MC estimate (3.0.1), a central performance question is how many samples $M$ are needed to make the approximation sufficiently accurate. To construct a reliable MC algorithm, it is important
to choose $M$ sufficiently large, but to make the algorithm efficient, it is important that it does not use more samples than needed to meet the accuracy demand; $M$ should also be low as possible. In this section we will demonstrate that the MC convergence rate is $\sigma M^{-1/2}$, and use this fact to construct a reasonable choice for $M$ in the setting of sampling Bernoulli r.v.

The vote of a citizen participating in a two party election might be modeled as a Bernoulli r.v. The American presidential election could be assigned the r.v. 

$$X(\omega) = \begin{cases} 1, & \text{if } \omega = \{\text{Republican}\} \\ 0, & \text{if } \omega = \{\text{Democrat}\}. \end{cases} \quad (3.1.1)$$

The quantity of interest is $\mu = P(X = 1)$; if $\mu < 0.5$, the Democrats win, and if $\mu > 0.5$, the Republicans win. Suppose we seek to predict the outcome of the election $\mu$ through sampling i.i.d. voters $X_1, X_2, \ldots$ by the sample average

$$\bar{X}_M := \frac{1}{M} \sum_{i=1}^{M} X_i,$$

with the accuracy-confidence constraint

$$P(|\bar{X}_M - \mu| > \text{TOL}) \leq \delta, \quad (3.1.2)$$

where we call TOL is the accuracy and $\delta$ the confidence. A reasonable question to raise is how large does $M$ have to be—how many voters must we sample—to achieve (3.1.2). For large $M$, the CLT motivates the approximation

$$P(|\bar{X}_M - \mu| > \text{TOL}) = P \left( \left| \frac{\sum_{i=1}^{M} X_i - \mu}{\sqrt{M} \sigma} \right| > \frac{\sqrt{M} \text{TOL}}{\sigma} \right) \approx 2 \left( 1 - \Phi \left( \frac{\sqrt{M} \text{TOL}}{\sigma} \right) \right).$$

implies that to ensure that

$$2 \left( 1 - \Phi \left( \frac{\sqrt{M} \text{TOL}}{\sigma} \right) \right) \leq \delta,$$

we should use

$$M = \frac{\sigma^2 \left( \Phi^{-1}(1 - \delta/2) \right)^2}{\text{TOL}^2} \quad (3.1.3)$$

vote samples. In general, the variance $\sigma^2$ is unknown, making the above expression for $M$ incomplete, but for Bernoulli r.v. $\sigma^2 = \mu(1 - \mu) \leq 1/4$, so we may conservatively choose $M$ according to

$$M = \frac{\left( \Phi^{-1}(1 - \delta/2) \right)^2}{4\text{TOL}^2}.$$
3.1. EXAMPLES AND PROPERTIES

So, if for example given the demands of accuracy $TOL = 0.02$ and confidence $\delta = 0.05$, we would, according to our derivations, have to sample at least $M = 2401$ i.i.d. voters.

The MC convergence rate $\sigma M^{-1/2}$ in the norm $\sqrt{E[|X_M - \mu|^2]}$ follows from noting that for i.i.d. samples

$$E\left[\frac{\sum_{i=1}^{M} X_i - \mu}{M^2}\right] = \frac{E[|X - \mu|^2]}{M} = \frac{\sigma^2}{M}.$$  \hspace{1cm} (3.1.4)

In euclidean norm, the convergence rate is often represented on the r.v. form

$$\frac{X_M - \mu}{\sqrt{M}} \sim \sigma \chi,$$

where $\chi$ denotes a standard normal r.v. and one should keep in mind that this is an asymptotic representation.

**Dimension Independent Convergence Rate**

To illustrate the dimension independent convergence rate of MC methods we consider an example of multivariate integration

$$I = \int_{A} g(x) \, dx.$$  \hspace{1cm} (3.1.5)

where $A$ is a compact subset of $\mathbb{R}^n$ and $g : A \rightarrow \mathbb{R}$. To approximate the value of $I$ by the MC method we generate i.i.d. r.v. $X_1, X_2, \ldots$ uniformly distributed in $A$ and set

$$\bar{I}_M = \sum_{i=1}^{M} \frac{g(X_i)}{M}.$$  \hspace{1cm} (3.1.6)

By construction we have that $E[\bar{I}_M] = I$ and by the reasoning of (3.1.4),

$$\sqrt{E[|\bar{I}_M - I|^2]} = \sqrt{Var(g(X))} M^{-1/2},$$

equaling the expected convergence rate. In comparison, a cubature\(^1\) approximation of (3.1.5) of order $k$ will have the convergence rate $\mathcal{O}(M^{-k/d})$ if we assume it resolves each dimension using $M^{1/d}$ grid points. So when $k/d < 1/2$, the MC method will asymptotically outperform the cubature method.

\(^1\)Cubature is the name for numerical integration in dimensions higher than 1.
To visualize the dimensional independent convergence we consider the problem of computing the volume of the unit ball in $\mathbb{R}^6$. This problem is equivalent to computing the integral (3.1.5) with

$$g(x) = \begin{cases} 1, & \text{if } |x| \leq 1, \\ 0, & \text{otherwise,} \end{cases}$$

and $A = [-1, 1]^n$. Figure 3.1 shows the performance of the MC algorithm compared to the first order cubature method

$$\tilde{I}_M := \sum_{i,j,k,l,m,n=1}^{[M^{1/6}]} g(x_i, x_j, x_k, x_l, x_m, x_n)h^6,$$  \hspace{1cm} (3.1.7)

where $[x] := \min\{n \in \mathbb{Z} | n \geq x\}$, $h = 2/[M^{1/6}]$, and $x_i = -1 + (i - 1/2)h$ for $i = 1, 2, \ldots, [M^{1/6}]$.

Figure 3.1: Comparison of convergence rates for approximating the volume of the unit ball in $\mathbb{R}^6$, $I = \pi^3/6$, when using either the MC algorithm (3.1.6) or the cubature (3.1.7). Due to the high dimensionality of the problem, the MC algorithm outperforms the first order cubature method. The observed convergence rates for the experiment fit quite well with theoretically predicted rates.
3.2 Variance Reduction

In Section 3.1, we observed that the convergence rate for the MC method is \( \sigma M^{-1/2} \) and that the number of samples needed to meet the accuracy-confidence constraint (3.1.2) was (approximately) given by

\[
M = \sigma^2 \left( \frac{\Phi^{-1}(1 - \delta/2)}{\text{TOL}} \right)^2.
\]

Generally, the cost of computing an MC estimate is

\[
O(M) = O\left( \sigma^2 \left( \frac{\Phi^{-1}(1 - \delta/2)}{\text{TOL}} \right)^2 \right),
\]

so if it somehow is possible to reduce the variance \( \sigma^2 \) of the samples used in the estimate, the computational cost of the MC estimate will also be reduced. Variance reduction is a vast subject for MC methods motivated from the goal of reducing computational cost. Here, we will restrict ourselves to describing the variance reduction technique named control variates which we will return to in Chapter 4. For a nice overview of other interesting variance techniques such as antithetic variables and importance sampling, we refer to [7].

Control Variates

Let \( X \) be a given r.v. which we seek to approximate the expected value \( E[X] \) by a Monte Carlo method, say

\[
\bar{X}_M = \frac{1}{M} \sum_{i=1}^{M} X_i.
\]

Suppose that \( Y \) is another r.v. for which the expected value \( E[Y] \) is known. The constructed random variable \( Z = X + E[Y] - Y \) will then have the same expected value as \( X \), so \( E[X] \) can also be approximated by applying the Monte Carlo method on the r.v. \( Z \), say

\[
\bar{Z}_M = \frac{1}{M} \sum_{i=1}^{M} Z_i.
\]

If the variance is reduced, \( \text{Var}(Z) = \text{Var}(X - Y) < \text{Var}(X) \), we reason from the introduction of this section that from a cost perspective, it is more efficient to approximate \( E[X] \) by sampling \( Z_i \) r.v. than by sampling \( X_i \) r.v.

3.3 Sequential Stopping Rules

Given no prior distributional information of samples used in an MC estimate, determining the number of samples \( M \) required to meet an accuracy-confidence constraint is very difficult; some distributional properties of the samples, such as their
variance, generally has to be inferred to choose $M$ sensibly. However daunting it might seem to perform reliable MC estimates in settings with no prior distributional sample information, such settings are frequently encountered and therefore they deserve some attention.

A sequential stopping rule MC algorithm progressively gathers distributional information on the samples through sampling higher moments and iteratively increasing the number of samples used in the estimate until a stop criterion is met. For example, Algorithm 1 represents a sample variance based stopping rule.

**Algorithm 1** Sample Variance Based Stopping Rule

**Input:** Initial number of samples $M_0$, accuracy TOL, confidence $1 - \delta$, and the CDF of the standard normal distributed r.v. $\Phi(x)$.

**Output:** $\bar{X}_M$.

Set $k = 0$, generate $M_k$ samples $\{X_i\}_{i=1}^{M_k}$ and compute the sample variance

$$\sigma^2_{M_k} := \frac{1}{M_k - 1} \sum_{i=1}^{M_k} (X_i - \bar{X}_{M_k})^2.$$  \hspace{1cm} (3.3.1)

**while** $2\left(1 - \Phi(\sqrt{M_k/TOL}/\sigma_{M_k})\right) > \delta$ **do**

Set $k = k + 1$ and $M_k = 2M_{k-1}$.

Generate a batch of $M_k$ i.i.d. samples $\{X_i\}_{i=1}^{M_k}$.

Compute the sample variance $\sigma^2_{M_k}$ as given in (3.3.1).

**end while**

Set $M = M_k$, generate samples $\{X_i\}_{i=1}^{M}$ and compute the output sample mean $\bar{X}_M$.

Chow and Robbins proved that for r.v. with bounded second moments, sample variance based stopping rules are asymptotically reliable, cf. [9]. That is, supposing the generalized sample average estimator $\sigma^2_M$ used is positive and asymptotically consistent in the sense $\sigma_M \to \sigma$ almost surely as $TOL \to 0$, then for any confidence $\delta \in (0, 1)$, the stopping criterion

$$\frac{\sigma^2_M}{M} \leq \frac{TOL^2}{(\Phi^{-1}(1 - \delta/2))^2}$$  \hspace{1cm} (3.3.2)

implies the asymptotic fulfillment of the accuracy-confidence constraint

$$\lim_{TOL \to 0} P(|\bar{X}_M - \mu| > TOL) \leq \delta.$$  

For practical purposes however, it is the reliability of sequential stopping rules for a fixed accuracy-confidence combination $TOL, \delta \in (0, 1)$ which is of interest; the non-asymptotic regime. To the best of our knowledge, this is largely an open problem. For a fixed number of samples $M$, an upper bound of the convergence rate in the
3.3. \textit{SEQUENTIAL STOPPING RULES}

non-asymptotic regime is given by the Berry-Esseen Theorem 2.2.8 which presents the rate $O(M^{-1/2})$. But, the leading order constant of the Berry-Esseen bound contains the factor

\[ \frac{\mathbb{E}[|X - \mu|^3]}{\sigma^2}, \]

which in the settings we are confined to would have to be sampled/estimated, giving a bound estimate instead of an explicit bound.

In Paper III of this thesis, we study stopping rules’ performance in the non-asymptotic regime. There we show by examples that for certain heavy-tailed r.v. and fixed $\text{TOL}, \delta \in (0, 1)$, the sample variance based stopping rule presented in Algorithm 1 fails to meet the accuracy-confidence constraint (3.1.2), and we construct a more reliable higher moments based stopping rule.
Chapter 4

Adaptive Weak Approximations for Stochastic Differential Equations

_It is not the strongest of the species that survives, nor the most intelligent that survives. It is the one that is the most adaptable to change._

—Charles Darwin

4.1 Stochastic Differential Equations (SDE)

SDE is an extension of ordinary differential equations that seeks to incorporate uncertainty into a differential equation:

\[
\frac{dX}{dt} = a(t, X)dt + \text{“noise”},
\]

\[X_0 = x_0,\]  \hspace{1cm} (4.1.1)

In general, the noise in the differential equation will be motivated from physical considerations for the model. If the “noise” is a Wiener process, the differential equation (4.1.1) becomes an SDE which (on Itô form) may be represented as follows

\[
\frac{dX}{dt} = a(t, X)dt + b(t, X)dW_t,
\]

\[X_0 = x_0.\]  \hspace{1cm} (4.1.2)

Here, the \(dW_t\) term is called the Wiener process increment.

The Wiener Process

The history of the Wiener process can be traced back to 1828 when the British botanist Robert Brown observed through a microscope that pollen grains suspended
in water performed an irregular motion. The physical process for the irregular motion was, in honor of its discoverer, given the name Brownian motion. The mathematician Norbert Wiener later proposed a mathematical model for standardized Brownian motion which was to become known as the Wiener process.

The Wiener process is a Gaussian process $W : [0, T] \times \Omega \rightarrow \mathbb{R}$ which is characterized by the properties

1. $W_0 = 0$,
2. The function $t \rightarrow W_t$ is almost surely continuous.
3. $W_t$ has independent increments with $W_t - W_s \sim \mathcal{N}(0, |t - s|)$.

As a consequence of property 3,

$$E[W_t W_s] = \min(t, s),$$

and, supposing $t_1 \leq t_2 \leq t_3 \leq t_4$,

$$E[(W_{t_2} - W_{t_1})(W_{t_4} - W_{t_3})] = E[W_{t_2} - W_{t_1}] E[W_{t_4} - W_{t_3}] = 0.$$

The Wiener process $W_t(\omega)$ is a function of two variables. For fixed $\omega \in \Omega$, we call $W_t(\omega) : [0, T] \rightarrow \mathbb{R}$ a realization or sample path of the Wiener process, cf. Figure 4.1. For each $t \in \mathbb{R}_+$ fixed, $W_t = W_t(\cdot)$ is a r.v. with the distribution (property 3) $W_t \sim \mathcal{N}(0, t)$, that is, for any Borel set $B \subset \mathbb{R}$,

$$P(W_t \in B) = \int_B \frac{\exp(-x^2/2t)}{\sqrt{2\pi t}} \, dx.$$

**Itô Integrals**

According to the theory for ordinary differential equations, we expect solutions of the SDE (4.1.2) to be representable on integral form

$$X_t = X_0 + \int_0^t a(s, X) \, ds + \int_0^t b(s, X) \, dW_s. \quad (4.1.3)$$

However, what that is meant by the rightmost integral integrating over Wiener increments $dW_s = W_{s+ds} - W_s$, is not clear. Let us investigate this issue by considering integrals on the form

$$I(f) = \int_0^T f(t, \omega) \, dW_t.$$
The theory of Riemann integrals for deterministic functions leads us to expect that the rightmost integral could be considered as the limit sum of infinitesimal contributions

$$\int_0^t f(t,\omega) \, dW_s = \lim_{\Delta t \to 0} \sum_i f(t_i^*,\omega) \Delta W_i$$  \hspace{1cm} (4.1.4)

with $\Delta W_i := W_{t_i+1} - W_{t_i}$ and $t_i^*$ any point in $[t_i, t_{i+1}]$. But, an example with $f(t,\omega) := W_t(\omega)$ will illustrate that the theory for deterministic integrals is not directly extendable to integrals involving stochastic processes: Choosing the leftmost point of each interval as integration points $t_i^* = t_i$ yields

$$I_-(f) := \lim_{\Delta t \to 0} \sum_i W_{t_i} \Delta W_i,$$

and, if instead choosing the rightmost point of each interval as integration points, $t_i^* = t_{i+1}$, we get

$$I_+(f) := \lim_{\Delta t \to 0} \sum_i W_{t_{i+1}} \Delta W_i.$$

From property 3. of the Wiener process, we may compute that while $\mathbb{E}[I_-(f)] = 0$, $\mathbb{E}[I_+(f)] = t$ on the other hand. This illustrates that for SDE it indeed does matter which point $t_j^* \in [t_i, t_{i+1}]$ of each infinitesimal interval is used in the integral evaluation (4.1.4).

For SDE, different choices of $t_j^*$ have given rise to different stochastic integrals, but in this thesis we will only consider Itô integrals. Itô integrals use the integrand
points \( t_j^* = t_j \), and is thus defined by

\[
\int_0^t f(s, \omega) \, dW_s := \lim_{\Delta t \to 0} \sum_i f(t_i, \omega) \Delta W_i.
\]

(4.1.5)

Itô integrals may be considered as the limiting form of the Forward Euler integrating method for SDE. In correspondence with Itô integrals, we call the SDE when written on the form (4.1.2) an Itô SDE. Note further that as an implication of having Wiener sample paths, the representation

\[
X_t(\omega) = X_0 + \int_0^t a(s, X_s(\omega)) \, ds + \int_0^t b(s, X_s(\omega)) \, dW_s(\omega),
\]

implies that for each fixed \( \omega, t \to X_t(\omega) \) is a sample path solution of the SDE, and for each fixed \( t, \omega \to X_t(\omega) \) is a r.v. See Figure 4.2 for an illustration of Itô SDE sample path solutions.

There are two types of solutions for Itô SDE; weak and strong solutions. In essence, one might say that weak solutions are sample path solutions which are "unique" in distributional sense, while strong solutions are unique in sample path sense according to the norm

\[
\|X\|_{L^2(P;[0,T])} = \mathbb{E} \left[ \int_0^T |X_t|^2 \, dt \right]^{1/2}
\]

See Section 4.2 for more on solution classification.

Itô Calculus

Itô calculus is a handy tool for analyzing Itô SDE. Let \( X_t \) be a solution of (4.1.2) and \( g(t, X_t) \) a sufficiently smooth function. Then the Itô "laws"

\[
dW_s^2 = dt, \quad dt \cdot dW_s = 0, \quad dt^2 = 0, \ldots
\]

(4.1.6)

yields following truncated Itô-Taylor expansion of \( dg \)

\[
dg(t, X_t) = \partial_t g(t, X_t) \, dt + \partial_X g(t, X_t) \, dX + \frac{1}{2} \partial_{XX} g(t, X_t) \, dX^2
\]

\[
= \left( \partial_t g(t, X_t) + a(t, X_t) \partial_X g(t, X_t) + \frac{(b(t, X_t))^2}{2} \partial_{XX} g(t, X_t) \right) \, dt
\]

\[
+ b(t, X_t) \partial_X g(t, X_t) \, dW_t.
\]

(4.1.7)

Here we used (4.1.2) and the Itô "laws" (4.1.6) to derive that \( dX^2 = b^2 dt + o(dt) \).
4.2. THE EULER METHOD FOR ITÔ SDE

Solving an SDE by Using Itô Calculus

For an application of Itô calculus, let us consider the following problem of geometric Brownian motion

\[ dX_t = rX_t dt + vX_t dW_t, \quad r, v \in \mathbb{R}_+. \]  \hspace{1cm} (4.1.8)

This problem might be interpreted as model for stock price evolution with \( r \) representing a constant interest rate, and \( v \) the constant volatility rate, i.e., the stochastic evolution of the stock price. To solve this SDE, we assume \( X > 0 \), and note that then (4.1.8) is equivalent to

\[ \frac{dX_t}{X_t} = r dt + vdW_t. \]

Considering the function \( g(X_t) = \log(X_t) \), we see by (4.1.7) with \( a(X_t) = rX_t \) and \( b(X_t) = vX_t \) that

\[ d\log(X_t) = \left( r - \frac{v^2}{2} \right) dt + vdW_t \implies X_t = X_0 \exp \left( \left( r - \frac{v^2}{2} \right) t + vW_t \right). \]

Figure 4.2 holds examples of sample path solutions of the SDE (4.1.8).

Figure 4.2: Left plot: Four independent sample path solutions of the SDE (4.1.8) when \( X_0 = 1, r = 0.1, \) and \( v = 0.05 \). Right plot: Four independent sample path solutions of the SDE (4.1.8) when \( X_0 = 1, r = 0.1, \) and \( v = 0.25 \).

4.2 The Euler Method for Itô SDE

We recall that a solution of an Itô SDE

\[ dX = a(t, X)dt + b(t, X)dW_t, \]
\[ X_0 = x_0. \]
can be written on the Itô integral form
\[ X_t = X_0 + \int_0^t a(s, X_s) \, ds + \int_0^t b(s, X_s) \, dW_s, \]
where the rightmost integral is the limit sum of a forward Euler integration
\[ \int_0^t b(s, X_s) \, dW_s := \lim_{\Delta t \to 0} \sum_i b(t_i, X_{t_i}) \Delta W_i. \]

For numerical solutions, the (forward) Euler method generates realizations \( \overline{X} \) of the Itô SDE by the scheme
\[
\overline{X}_{t_{n+1}} = \overline{X}_{t_n} + a(t_n, \overline{X}_{t_n}) \Delta t + b(t_n, \overline{X}_{t_n}) \Delta W_n,
\]
with uniform step size \( \Delta t = T/N \), \( t_n = n \Delta t \), and Wiener increments \( \Delta W_n = W(t_{n+1}) - W(t_n) \sim N(0, \Delta t) \).

The rate of convergence of \( \overline{X}_T \) to the real solution \( X_T \) depends on the measure. The strong convergence is given by \( \mathbb{E}[|X_T - \overline{X}_T|] \) and is a measure for the pathwise convergence of each numerical realization \( \overline{X}_T(\omega) \) to its corresponding real solution realization \( X_T(\omega) \). For many problems, however, pathwise convergence might be more than you seek. Instead, you might be interested in some distributional quantity of interest expressible on the form \( \mathbb{E}[g(X_t)] \) for some smooth function \( g \). If so, the convergence criterion of interest is the weak convergence\(^2\)
\[
\sup_{g \in C_c^\infty(\mathbb{R})} |\mathbb{E}[g(X_T)] - \mathbb{E}[g(\overline{X}_T)]|,
\]
where \( C_c^\infty(\mathbb{R}) \) denotes the set of smooth compactly supported functions. For the Euler method on uniform grids with step size \( \Delta t \), the strong convergence rate is \( O(\Delta t^{1/2}) \), cf. [22, 8], and the weak convergence was shown by Talay and Tubaro to be \( O(\Delta t) \), cf. [31], see also cf. [8].

Remark 4.2.1 In this thesis we consider SDE solved numerically using the Euler method, but we remark that through truncation of Itô-Taylor expansions, one may develop higher order numerical schemes for SDE. One example is the Milstein scheme,
\[
\overline{X}_{n+1} = \overline{X}_n + a_n \Delta t + b_n \Delta W_n + \frac{b_n \partial_X b_n}{2} (\Delta W_n^2 - \Delta t),
\]
which has first order strong convergence. As for deterministic problems, however, higher order methods require stronger regularity assumptions and are generally more difficult to expand to adaptive methods in higher dimensions. For more on higher order methods, we refer to [22].

\[^2\]Weak convergence is equivalent to convergence in distribution, as defined in Chapter 2.
An adaptive Euler Method for Weak Solutions

For problems where either the drift or diffusion in the SDE lacks regularity, numerical solutions using adaptive time stepping may improve convergence rates, or even be your only option for obtaining reliable numerical solutions. In this subsection, we will describe the adaptive Euler method for weak solutions of SDE which was developed by Szepessy et al. in [30] in the beginning of the 2000s.

Objective

Consider the problem of generating solutions of the initial value SDE

\[
dX = a(t, X)dt + b(t, X)dW_t, \quad 0 \leq t \leq T, \\
X_0 = x_0, \quad x_0 \in \mathbb{R}. 
\]  

with an adaptive Euler method which fulfills the weak accuracy constraint

\[
\left| E[g(X_T) - g(\bar{X}_T)] \right| \leq \text{TOL}_T, \tag{4.2.3}
\]

where \(X_t\) and \(\bar{X}_t\) represents the explicit and numerical solution of (4.3.1), respectively, and \(\text{TOL}_T > 0\) is a given tolerance constraint. In this setting, the goal of adaptivity is to minimize \(N\), the number of adaptive time steps needed in the grid \(0 = t_0 < t_1 < \ldots < t_N = T\) such that your adaptive Euler solution \(\bar{X}_t\) fulfill (4.2.3).

As expected, the adaptive Euler method is given by

\[
\bar{X}_{t_{n+1}} = \bar{X}_{t_n} + a(t_n, \bar{X}_{t_n})\Delta t_n + b(t_n, \bar{X}_{t_n})\Delta W_n, 
\]

where \(\Delta t_n(\omega) = (t_{n+1} - t_n)(\omega)\) denote the \(n\)-th time step of the sample path solution \(\omega\).

Grid Refinement

The adaptive method by Szepessy et al. expresses the weak discretization error by the error expansion

\[
E[g(X_T) - g(\bar{X}_T)] \simeq \sum_{n=0}^{N} E[\rho(t_n, \cdot)\Delta t_n^2(\cdot)] + \text{h.o.t.} \tag{4.2.5}
\]

Here, \(\rho(t_n, \omega)\) represents the error density for a given sample path solution \(\omega\) at the time \(t_n\) and the error indicators \(\rho(t_n, \omega)\Delta t_n^2(\omega)\) then represents, up to higher order terms, the error contribution from the \(n\)-th time step of the sample path \(\omega\). The error indicators \(\rho(t_n, \omega)\Delta t_n^2(\omega)\) provide information for further refinement of the time grid. Error control at a low computational cost is obtained through the condition

\[
\rho(t_n, \omega)\Delta t_n^2(\omega) \leq C^{\text{TOL}_T-E[N]}, \quad \forall n = 0, 1, 2, \ldots, N(\omega). \tag{4.2.6}
\]
Given an initial time grid, time increments, and a Wiener process numerical sample path which we represent by \( t.(\omega, 0) \), \( \Delta t.(\omega, 0) \), and \( W_t.(\omega, 0) \), respectively, the following adaptive mesh refinement algorithm is used by Szepessy et al. to ensure that the condition (4.2.6) is met.

1. Set \( \ell = 0 \).
2. Compute the error density \( \rho(t., \omega) \) for the numerical path represented by \( t.(\omega, \ell) \), \( \Delta t.(\omega, \ell) \), and \( W_t.(\omega, \ell) \), cf. (4.2.9).
3. If condition (4.2.6) is fulfilled, then stop; otherwise
4. Refine the grid \( t.(\omega, \ell) \) at all points where (4.2.6) is not fulfilled by halving the steps:

\[
\Delta t_{\bar{n}}(\omega, \ell + 1) = \frac{\Delta t_n(\omega, \ell)}{2} \quad \text{and} \quad \Delta t_{\bar{n}+1}(\omega, \ell + 1) = \frac{\Delta t_n(\omega, \ell)}{2}, \quad (4.2.7)
\]

where \( \bar{n} \geq n \) is defined as the natural number such that \( t_{\bar{n}}(\omega, \ell + 1) = t_n(\omega, \ell) \).

When adding a point \( t_{\bar{n}}(\omega, \ell + 1) \) to your grid by the refinement (4.2.7), add a corresponding Wiener process sample path point by Brownian bridges:

\[
W_{t_{n+1}}(\omega, \ell + 1) = W_{t_n}(\omega, \ell) + \frac{\sqrt{\Delta t_n(\omega, \ell)}}{2} \xi, \quad \text{where} \quad \xi \sim \mathcal{N}(0, 1).
\]

5. Set \( \ell = \ell + 1 \) and return to 2.

**The Error Density**

For notational convenience, let us write the Euler method scheme (4.2.4) as follows

\[
X_{t_{n+1}} = c(t_n, X_{t_n}), \quad \text{where} \quad c(t_n, x) := x + a(t_n, x) \Delta t_n + b(t_n, X_{t_n}) \Delta W_n. \quad (4.2.8)
\]

The error density derived in Theorem 2.2 of [30] can then be represented by

\[
\rho(t_n, \cdot) = \frac{1}{2\Delta t_n} \left( (a(t_{n+1}, \bar{X}_{t_{n+1}}) - a(t_n, X_{t_n})) \phi(t_{n+1}) + (b^2(t_{n+1}, \bar{X}_{t_{n+1}}) - b^2(t_n, X_{t_n})) \partial_{X_{t_n}} \phi(t_{n+1}) \right), \quad (4.2.9)
\]

where \( \phi \) is the solution of the discrete dual backward problem

\[
\phi(t_n) = \partial_x c(t_{n+1}, \bar{X}_{t_{n+1}}) \phi(t_{n+1}), \quad t < T
\]

\[
\phi(T) = g'(\bar{X}_T), \quad (4.2.10)
\]

and \( \partial_{X_{t_n}} \phi(t_n) \) is obtained by a longer scheme by linearising the forward problem (4.1.2), cf. [30, Theorem 2.2].
4.2 THE EULER METHOD FOR İTO SDE

To derive the error density (4.2.9), we extend the numerical solutions of the Euler method (4.2.4) to \( t \in [t_n, t_{n+1}] \) by

\[
\bar{X}_t = \bar{X}_{t_n} + \int_{t_n}^t \bar{a}(s; \bar{X}) \, ds + \int_{t_n}^t \bar{b}(s; \bar{X}) \, dW_s,
\]

where \( \bar{a} \) and \( \bar{b} \) are piecewise constant approximations

\[
\bar{a}(s; \bar{X}) = a(t_n, X_{t_n}) \quad \text{and} \quad \bar{b}(s; \bar{X}) = b(t_n, X_{t_n}) \quad \text{for} \quad s \in [t_n, t_{n+1}], \quad n = 0, 1, 2, \ldots
\]

The expected value \( \mathbb{E}[g(X_T)] \) with \( X_T \) solving the SDE (4.1.2) is related to a Partial Differential Equation (PDE): the utility function \( u(t, x) := \mathbb{E}[g(X_T)|X_t = x] \) solves the Kolmogorov backward equation

\[
\begin{align*}
\partial_t u &= - \left( a \partial_x + \frac{b^2}{2} \partial_{xx} \right) u, \quad (t, x) \in [0, T] \times \mathbb{R} \quad (4.2.11) \\
u(T, x) &= g(x),
\end{align*}
\]

cf. [22]. By Itô calculus and the relation (4.2.11), it follows that

\[
du(t, X_t) = \left( \partial_t u(t, X_t) + \bar{a}(t; X) \partial_x u(t, X_t) + \frac{\bar{b}^2(t; X)}{2} \partial_{xx} u(t, X_t) \right) dt \\
+ \bar{b}(t; X) \partial_x u(t, X_t) \, dW_t
\]

\[
= \left( (a(t, X_t) - a(t, X_t)) \partial_x u(t, X_t) + \frac{b^2(t, X_t) - b^2(t, X_t)}{2} \partial_{xx} u(t, X_t) \right) dt \\
+ b(t; X) \partial_x u(t, X_t) \, dW_t.
\]

Recalling that \( X_0 = X_0 \), we see that \( u(0, X_0) = \mathbb{E}[g(X_T)] \). Hence,

\[
\begin{align*}
\mathbb{E}[g(X_T)] - g(X_T) &= - \int_0^T du(t, X_t) \\
= & \sum_{n} \int_{t_n}^{t_{n+1}} \left( a(t, X_t) - a(t, X_{t_n}) \right) \partial_x u(t, X_t) + \frac{b^2(t, X_t) - b^2(t, X_{t_n})}{2} \partial_{xx} u(t, X_t) \, dt \\
- & \sum_{n} \int_{t_n}^{t_{n+1}} b(t_n, X_{t_n}) \partial_x u(t, X_t) \, dW_t.
\end{align*}
\]

(4.2.12)

Taking the expected value of equation (4.2.12) yields

\[
\begin{align*}
\mathbb{E}[g(X_T) - g(X_T)] &= \sum_{n} \int_{t_n}^{t_{n+1}} \mathbb{E} \left[ \left( a(t, X_t) - a(t, X_{t_n}) \right) \partial_x u(t, X_t) \right] \, dt \\
+ & \sum_{n} \int_{t_n}^{t_{n+1}} \mathbb{E} \left[ \frac{b^2(t, X_t) - b^2(t, X_{t_n})}{2} \partial_{xx} u(t, X_t) \right] \, dt,
\end{align*}
\]

(4.2.13)
where we used that the expected value of the integral with $dW_t$ increments in (4.2.12) is zero, cf. [28]. Introducing the utility function for the numerical solution $\bar{u}(t, x) := E\left[g(X_T) | X_t = x\right]$, the following approximations of the terms in (4.2.13) are valid:

\[
\int_{t_n}^{t_{n+1}} E\left[a(t, X_t) - (a(t_n, X_{t_n}))\partial_x u(t, X_t)\right] dt = E\left[(a(t_{n+1}, X_{t_{n+1}}) - a(t_n, X_{t_n}))\partial_x \bar{u}(t_{n+1}, X_{t_{n+1}})\right] \frac{\Delta t_n}{2} + \text{h.o.t.}
\]

and

\[
\int_{t_n}^{t_{n+1}} E\left[\frac{b^2(t, X_t) - b^2(t_n, X_{t_n})}{2} \partial_{xx} u(t, X_t)\right] dt = E\left[\frac{b^2(t_{n+1}, X_{t_{n+1}}) - b^2(t_n, X_{t_n})}{2} \partial_{xx} \bar{u}(t_{n+1}, X_{t_{n+1}})\right] \frac{\Delta t_n}{2} + \text{h.o.t.}
\]

Further, since

\[
\partial_x \bar{u}(t, x) = \frac{\partial}{\partial x} E\left[g(X_T) | X_t = x\right] = E\left[g'(X_T) \partial_X X_T | X_t = x\right],
\]

we see that

\[
\partial_x \bar{u}(t_{n+1}, X_{t_{n+1}}) = E\left[g'(X_T) \partial_{X_{t_{n+1}}} X_T | \mathcal{F}_{t_{n+1}}\right], \quad (4.2.14)
\]

where $\mathcal{F}_t$ denotes the $\sigma$-algebra generated by $\{W_s\}_{s \in [0, t]}$. By definition (4.2.10),

\[
0 = \sum_{\ell=n}^{N-1} (\phi(t_{\ell+1}) - \phi(t_\ell)) \partial_{X_{t_\ell}} X_{t_{\ell+1}} = \sum_{\ell=n}^{N-1} \phi(t_{\ell+1}) \left(\partial_{X_{t_{\ell+1}}} X_{t_{\ell+1}} - \partial_{X_{t_\ell}} c(X_{t_\ell}, t) \partial_{X_{t_\ell}} X_t\right) - \phi(T) \frac{\partial X_T}{\partial X_{t_n}} + \phi(t_n) \frac{\partial X_{t_n}}{\partial X_{t_n}}
\]

\[
= -\phi(T) \frac{\partial X_T}{\partial X_{t_n}} + \phi(t_n).
\]

Recalling that $\phi(T) = g'(X_T)$, we see that $\phi(t_n) = g'(X_T) \partial_{X_{t_n}} X_T$. Since $a(t_{n+1}, X_{t_{n+1}}) - a(t_n, X_{t_n})$ is $\mathcal{F}_{t_{n+1}}$ measurable and by (4.2.14), we derive that

\[
E\left[(a(t_{n+1}, X_{t_{n+1}}) - a(t_n, X_{t_n}))\partial_x \bar{u}(t_{n+1}, X_{t_{n+1}})\right] = E\left[(a(t_{n+1}, X_{t_{n+1}}) - a(t_n, X_{t_n}))E[\phi(t_{n+1}) | \mathcal{F}_{t_{n+1}}]\right] = E\left[(a(t_{n+1}, X_{t_{n+1}}) - a(t_n, X_{t_n}))\phi(t_{n+1})\right].
\]

(4.2.16)
and by a similar argument it follows that
\[
\begin{align*}
E \left[ b^2(t_{n+1}, X_{t_{n+1}}) - b^2(t_n, X_{t_n}) \frac{\partial_{xx} u(t_{n+1}, X_{t_{n+1}})}{2} \right] \\
= E \left[ b^2(t_{n+1}, X_{t_{n+1}}) - b^2(t_n, X_{t_n}) \frac{\partial_{xx} \phi(t_{n+1})}{X_{t_{n+1}}} \right].
\end{align*}
\]
(4.2.17)

We conclude the derivation of the density representation (4.2.9) by inserting (4.2.16) and (4.2.17) into equality (4.2.13).

**Remark 4.2.2** By a longer analysis, Szepessy et al. refines the error density to the following representation
\[
\begin{align*}
\rho(t_n, \omega) &= \frac{1}{2} \left( \partial_t a + a \partial_x a + \frac{b^2 \partial_{xx} a}{2} \right) \phi(t_{n+1}) \\
&+ \left( \partial_t b^2 + a \partial_x b^2 + \frac{b^2 \partial_{xx} b^2}{2} + b^2 \partial_x a \right) \frac{\partial_{xx} \phi(t_{n+1})}{X_{t_{n+1}}} \\
&+ \frac{b^2}{2} \frac{\partial_{xx} b^2 \partial^2 \phi(t_{n+1})}{X^2_{t_{n+1}}},
\end{align*}
\]
cf. Theorem 3.3 of [30].

**Remark 4.2.3** The use of dual functions is standard in both in optimal control theory and in adaptive grid algorithms for ordinary and partial differential equations, cf. [1, 13].

### 4.3 Adaptive Weak Solution Approximation

For the Itô SDE
\[
\begin{align*}
\frac{dX}{dt} &= \frac{a(t, X) dt + b(t, X) dW_t, \quad 0 \leq t \leq T,}{} \\
X_0 &= x,
\end{align*}
\]
we consider the problem of minimizing the computational cost of approximating \(E[g(X(T))]\) within a given tolerance \(TOL > 0\). Problems of this kind arise, for example, in computing option prices in mathematical finance, cf. [21] and [17].

Approximations of weak solutions are typically obtained using Monte Carlo (MC) methods. For SDE problems, it is possible to reduce the variance of the MC estimate, and thus reduce the complexity, by generating solution realizations on grids of different step size. This variance reduction technique is called Multilevel Monte Carlo (MLMC) methods and is in some ways in some ways similar to multigrid methods for PDE problems [6]. Paper II of this thesis develops an adaptive time step MLMC algorithm. As a preparation for that paper, we will here give an outline of the uniform time step MLMC algorithm. But let us first consider the single level MC method.
CHAPTER 4. ADAPTIVE WEAK APPROXIMATIONS FOR SDE

Single Level Weak Approximation

The single level MC method generates $M$ numerical realizations $\bar{X}_T(\omega_i)$ by the uniform time step Euler method and approximates $\mathbb{E}[g(X_T)]$ by the MC sample average

$$A(g(\bar{X}_T); M) = \sum_{i=1}^{M} \frac{g(\bar{X}_T(\omega_i))}{M}. \quad (4.3.2)$$

According to the problem formulation, we seek to fulfill the weak error bound $|\mathbb{E}[g(X_T)] - A(g(\bar{X}_T); M)| \leq \text{TOL}$ at minimal computational cost. For error control, the approximation error is split into two parts

$$|\mathbb{E}[g(X_T)] - A(g(\bar{X}_T); M)| \leq |E[g(X_T)] - \mathbb{E}[g(\bar{X}_T)]|$$
$$+ |E[g(X(T))] - A(g(\bar{X}_T); M)| =: \mathcal{E}_T + \mathcal{E}_S,$$

where $\mathcal{E}_T$ is the time discretization error and $\mathcal{E}_S$ the statistical error. Under appropriate smoothness assumptions on the drift $a$ and diffusion $b$, we mentioned in Section 4.2 that weak convergence rate, i.e., the time discretization error, for the Euler method fulfills $\mathcal{E}_T = O(\Delta t)$. Further, by the CLT, the asymptotic convergence rate for the statistical error is $\mathcal{E}_S = O(M^{-1/2})$. So to ensure that $|E[g(X(T))] - A(g(\bar{X}_T); M)| = O(\text{TOL})$, one should choose $\Delta t = O(\text{TOL})$ and $M = O(\text{TOL}^{-2})$. The computational cost thus becomes $O(\text{TOL}^{-3})$ for the single level MC method.

The Multilevel Monte Carlo Method

The MLMC method developed by Giles in [16] expanded the single level MC method by constructing Euler method realizations of the SDE (4.3.1) on hierarchies of uniform time grids, typically with the size relations

$$\Delta t^{(\ell)} = 2^{-\ell} \Delta t^{(0)}, \quad \ell = 0, 1, 2, \ldots, L.$$

Let $\bar{X}^{(\ell)}_t$ denote an Euler method realization on a grid with uniform step size $\Delta t^{(\ell)}$. Then the MLMC method approximates $\mathbb{E}[g(X(T))]$ by the telescoping sum

$$A_{\text{MC}}(g(\bar{X}_T); M_0) = \sum_{i=1}^{M_0} \frac{g(\bar{X}_T^{(0)}(\omega_{i,0}))}{M_0} + \sum_{\ell=1}^{L} \sum_{i=1}^{M_{\ell}} g(\bar{X}_T^{(\ell)}(\omega_{i,\ell})) - g(\bar{X}_T^{(\ell-1)}(\omega_{i,\ell})) \frac{M_{\ell}}{M_0}, \quad (4.3.3)$$

where $M_0$ and $M_\ell := 2^{-\ell} M_0$, $\ell = 1, \ldots, L$, represents the number of samples generated at respective grid levels. On each level $\ell$ in the above estimator, the realization pairs $\bar{X}_T^{(\ell)}(\omega_{i,\ell})$ and $\bar{X}_T^{(\ell-1)}(\omega_{i,\ell})$ are generated by the same Wiener process sample path $W_t(\omega_{i,\ell})$, but on different temporal grids with step size $\Delta t^{(\ell)}$ and $\Delta t^{(\ell-1)}$, respectively. The values of a Wiener process sample path is first computed on the
coarse grid $t_n^{(\ell-1)} = n\Delta t^{(\ell-1)}$, let us write $W_{t_n^{(\ell-1)}}(\omega)$, and thereafter, the values of the sample path on the finer grid $t_n^{(\ell)} = n\Delta t^{(\ell)}$ is computed by Brownian bridges

$$W_{t_n^{(\ell)}}(\omega) = W_{t_n^{(\ell-1)}}(\omega) + \frac{W_{t_n^{(\ell-1)}}(\omega) + W_{t_n^{(\ell-1)}}(\omega)}{2} + \frac{\sqrt{\Delta t^{(\ell-1)}}}{2} \xi,$$

where $\xi \sim \mathcal{N}(0, 1)$, cf. Figure 4.3.

Figure 4.3: **Left plot:** A sample path $W_t(\omega)$ plotted on the coarse grid $W_{t_n^{(0)}}(\omega)$ (blue line) with $\Delta t^{(0)} = 1/10$, and its finer level pair generated by Brownian bridges, $W_{t_n^{(1)}}(\omega)$ (green line) with $\Delta t^{(1)} = \Delta t^{(0)}/2$. **Right plot:** Euler method numerical solutions of the Ornstein-Uhlenbeck SDE problem $dX_t = 2(1 - X_t)dt + 0.2dW_t$, $X_0 = 3/2$ for the Wiener process sample path plotted in the left plot. $\overline{X}_{t_n^{(0)}}(\omega)$ (blue line) generated from using the Wiener increments from the path $W_{t_n^{(0)}}(\omega)$ and $\overline{X}_{t_n^{(1)}}(\omega)$ (green line) generated using Wiener increments from the path $W_{t_n^{(1)}}(\omega)$

The MLMC method has the consistent estimator

$$E[A_L(g(\overline{X}_T); M_0)] = E[g(\overline{X}_T^{(L)})],$$

and since sample paths are i.i.d.

$$E\left[\left(g(\overline{X}_T^{(\ell)}) - g(\overline{X}_T^{(\ell-1)})\right)\left(g(\overline{X}_T^{(m)}) - g(\overline{X}_T^{(m-1)})\right)\right] = 0, \text{ when } \ell \neq m.$$

The strong convergence $E\left[\left(g(\overline{X}_T^{(\ell)}) - g(X_T)\right)^2\right] = O(\Delta t^{(\ell)})$, cf. Section 4.2,
further implies that
\[
\text{Var}(A_{\text{MLC}}(g(X_T); M_0)) = \frac{\text{Var}(g(X_T^{(0)}))}{M_0} + \sum_{\ell=1}^{L} \frac{\text{Var}(g(X_T^{(\ell)}) - g(X_T^{(\ell-1)}))}{M_\ell}
\]

\[= \mathcal{O}\left(M_0^{-1} + \sum_{\ell=1}^{L} 2^{-\ell}M_\ell^{-1}\right) = \mathcal{O}(LM_0^{-1}). \tag{4.3.4}\]

Giles showed that by choosing \( L = \mathcal{O}(\log(\text{TOL}^{-1})) \) and \( M_0 = \mathcal{O}(\text{LTOL}^{-2}) \), one obtains

\[ |E[g(X_T)] - A_{\text{MLC}}(g(X_T); M_0)| \leq \mathcal{O}(\text{TOL}) \]

at the computational cost

\[ \mathcal{O}\left(\sum_{\ell=0}^{L} \frac{M_\ell}{\Delta t^{(\ell)}}\right) = \mathcal{O}\left((\log(\text{TOL}^{-1}))^2 \text{TOL}^{-2}\right). \]

This might be shown by splitting MLMC error into two contributions

\[ |E[g(X_T)] - A_{\text{MLC}}(g(X_T); M_0)| \leq \left|E\left[g(X_T) - g(X_T^{(L)})\right]\right| + \left|E\left[g\left(X_T^{(L)}\right) - A_{\text{MLC}}(g(X_T); M_0)\right]\right| =: \mathcal{E}_T + \mathcal{E}_S. \]

When \( L = \mathcal{O}(\log(\text{TOL}^{-1})) \) and \( M_0 = \mathcal{O}(\text{LTOL}^{-2}) \), the weak rate of convergence for the Euler method implies that \( \mathcal{E}_T = \mathcal{O}(\text{TOL}) \), and by the CLT and (4.3.4), one may derive that \( \mathcal{E}_S = \mathcal{O}(\text{TOL}) \).

With Giles’ MLMC method, the computational cost is thus improved from \( \mathcal{O}(\text{TOL}^{-3}) \) for the single level MC method to \( \mathcal{O}\left((\log(\text{TOL}^{-1}))^2 \text{TOL}^{-2}\right) \). The cost improvement is due to the multilevel variance reduction which is similar to control variates, cf. Section 3.2. That is, the multilevel average operator \( A_{\text{MLC}}(g(X_T); M_0) \) has lower variance than the single level average operator \( A(g(X_T); M) \), and this implies that fewer samples have to be used to control the statistical error for the MLMC method than for the single level MC method.

**Remark 4.3.1** The theory presented for 1-dimensional SDE problems in this chapter generalizes to the \( n \)-dimensional setting, cf. [8, 22, 16]. MC methods for weak approximations of SDE do in fact become more interesting in higher dimensions due to the dimension independent convergence rate of MC methods, cf. Section 3.1. Lower dimensional weak approximation problems are often more efficiently solved by solving the Kolmogorov backward partial differential equation relating to the utility function \( u(t,x) = E[X_T|X_t = x] \), cf. (4.2.11).
Chapter 5

Wireless Channel Modeling

Wireless signal transmission is realized through electromagnetic radiation from transmitter to receiver, and the signal received is described by electromagnetic field at the receiver as a function of time. In principle, a numerical solution of Maxwell’s equations with well resolved scattering boundary would yield the electrical field at the receiver and everywhere around it. But, for standard Modeling scenarios where the communication carrier frequency is of order $10^9$Hz and the wavelength consequently is of order centimeters, localization of the scattering boundary would have to be made to the order of centimeter accuracy to obtain an acceptable electrical field solution. Since this typically implies the need for resolving billions of boundary points, determining the electrical field through solving Maxwell’s equations is considered too costly both from a boundary measurement and computational perspective.

Multipath Fading Channel (MFC) models give a more cost efficient, but less accurate model of the received signal than Maxwell’s equations by superpositioning a large number of incoming signal wave paths. In Paper I of this thesis, we study an MFC model with noise introduced by scatterers flipping on and off. To prepare the reader for the contents that paper, we will here give a short introduction to MFC models and some more general signal theory concepts.

5.1 The Multipath Fading Channel

An MFC model approximates the output signal of a wireless channel by a superposition of a large number $M$ of incoming signal wave paths, cf. Figure 5.1. With $X_t$ denoting the baseband input signal, the time invariant MFC model has the baseband output representation

$$Z_{t,M} = \sum_{j=1}^{M} a(\alpha_j)e^{-i2\pi f c \tau(\alpha_j,t)} X_{t-\tau(\alpha_j,t)},$$

(5.1.1)
where $f_c$ denotes the carrier frequency, $\alpha_j$ the horizontal angle of arrival for the $j^{th}$ wave path, and $a(\alpha_j)$ and $\tau(\alpha_j, t)$ its amplitude and time delay function, respectively.

The terminology baseband and passband refers to the frequency modulation of the signal in question. Let $\mathcal{F}(\cdot)$ denote the Fourier transform. Then the baseband input signal $X_t$ has the spectral representation $\mathcal{F}(X_t)(f) \in C([-W,W])$. The spectral representation is centered at origo, band-limited to $[-W,W]$, and it does not have to be symmetric w.r.t. $f$; the baseband signal $X_t$ may be complex-valued.

Before transmission the baseband signal $X_t$ is modulated to the carrier frequency $f_c$ by the procedure

$$\tilde{X}_t = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(X_t)(f - f_c) + \mathcal{F}(X_t)(-f - f_c)}{\sqrt{2}}\right)(t),$$

or, in time domain, $\tilde{X}_t = \sqrt{2}\text{Re}[X_t \exp(i2\pi f_c t)]$. The transmitter transmits the real-valued passband input signal $\tilde{X}_t$, and the receiver receives the passband output signal

$$\tilde{Z}_t = \int_{\mathbb{R}} \tilde{H}(s,t)\tilde{X}_{t-s} \, ds \quad (5.1.2)$$
where the passband channel response for MFC channels is given by

\[ \tilde{H}(s, t) := \sum_{j=1}^{M} a(\alpha_j) \delta(s - \tau(\alpha_j, t)), \quad (5.1.3) \]

and \( \delta(\cdot) \) denotes the Dirac delta function. To extract the information of the received signal, the passband output signal is demodulated to the baseband output signal

\[ Z_t = \mathcal{F}^{-1} \left( \sqrt{2} \mathcal{F}(\tilde{Z}_\cdot)(f - f_c) \right)_{|f| < W}(t). \]

Frequency modulation makes it possible to operate many channels simultaneously in the same physical area by transmitting at different carrier frequencies (for example, FM radio broadcasting with many active channels), but the information each channel transmits is contained and most simply analyzed as a baseband signal. In fact, by introducing the baseband channel response

\[ H(s, t) := \sum_{j=1}^{M} a(\alpha_j) e^{-i2\pi f_c \tau(\alpha_j, t)} \delta(s - \tau(\alpha_j, t)), \quad (5.1.4) \]
it is possible to represent the output baseband as a function of the input baseband

\[ Z_t = \int_{\mathbb{R}} H(s,t)X_{t-s} \, ds, \quad (5.1.5) \]

cf. [32]. This representation equals (5.1.1), avoids dealing with the passband, and it shows that given the baseband input signal, the output signal is described by the baseband channel response.

In Paper I, we study the setting with baseband input signal \( X_t := 1 \). This setting gives the baseband output signal

\[ Z_{t,m} = \sum_{j=1}^{M} a(\alpha_j)e^{-i2\pi f_c \tau(\alpha_j, t_m)}, \]

which is closely related to the discrete channel response.

Although the wireless channel (5.1.1) is presented in a continuum setting, real life channels signal processing is performed on discrete time steps with all wave paths arriving at the receiver within a sample period \( \Delta t \) are averaged to become the received value for that time sample, cf. [32]. In the transition from continuum to discrete, the Nyquist-Shannon Sampling Theorem describes how small the sample period \( \Delta t \) have to be to resolve a continuous signal \( X_t \).

**Theorem 5.1.1 (Nyquist-Shannon Sampling Theorem [33])** Suppose a function \( X_t \) contains no frequencies higher than \( B \) Hertz. Then it is completely determined by giving its ordinates at a series of points spaced \( 1/(2B) \) seconds apart.

**Remark 5.1.2** For more on the ideas presented in this wireless channels, see [32] and [11].

**Clarke’s Model**

Clarke’s model is a famous MFC model which was motivational for our work, cf. [10]. It considers the superposition of \( M \) wave paths with amplitudes \( a = 1/\sqrt{M} \);

\[ Z_{t,M} = \frac{1}{\sqrt{M}} \sum_{m=1}^{M} e^{-i(2\pi f_c v \cos(\alpha_m) t/c + \theta_m) \phi}, \quad (5.1.6) \]

Here, apart from the variables already introduced in Section 5.1, \( c \) is the speed of light, \( \{\theta_m\}_{m=1}^{M} \) are i.i.d. initial phase shifts uniformly distributed in \([0, 2\pi]\) and the scatterer angle of arrivals \( \{\alpha_m\}_{m=1}^{M} \) are distributed according to the scatterer angle density \( p(\alpha) \) which is independent from the initial phase shift distribution. The delay function is on the form \( \tau(\alpha, t) = -f_c v \cos(\alpha) t/c \) and the factor \( \partial_t \tau(\alpha, t) = -f_c v \cos(\alpha)/c \) is the Doppler shift under the assumption that the receiver moves in the direction \((1, 0)\). The Doppler shift describes the change of frequency of a wave
for a receiver moving relative to the wave transmitter (the tone of the siren of a passing ambulance is a classical illustration of this phenomenon).

Among the functions used to analyze channel properties is the autocorrelation function and the Power Spectral Density (PSD). The autocorrelation function $A_M(t) := E[Z_{t,M}Z_{0,M}^*]$ describes the correlation between $Z_{t,M}$ and $Z_{0,M}$, and for Wide Sense Stationary (WSS) signals, the PSD is the Fourier transform pair of the autocorrelation function. Let us state this formally.

**Definition 5.1.3 (Wide Sense Stationary Random Process)** A random process $Z_t$ is WSS if there is a function $g$ such that

$$E[Z_{\tilde{t},M}Z_{t,M}^*] = g(|\tilde{t} - t|) \quad \text{for any } \tilde{t}, t \in \mathbb{R}.$$

**Theorem 5.1.4 (Wiener-Khintchine Theorem [11, p. 49])** The power spectral density and autocorrelation of a wide sense stationary process are Fourier transform pairs.

Considering the scenario with scatterer angle density $p(\alpha) = \frac{1}{2\pi}$, Clarke noted that for his model the autocorrelation function $A_M(t)$ converges to the zeroth-order Bessel function of the first kind, $J_0(2\pi f_c v t / c)$, as $M \to \infty$, and that its PSD then is on the form

$$S(f) = \begin{cases} \frac{c}{\pi \sqrt{(v f_c)^2 - (v f_c)^2}} & |f| < v f_c / c \\ 0 & |f| \geq v f_c / c, \end{cases}$$

(5.1.7)

cf. Figure 5.3.

**MFC Model with Flipping Scatterers**

Due to local shadowing by moving cars, pedestrians, leaves blowing in the wind, weather conditions etc., scatterers can flip from being active to passive and vice versa. Seeking to include scattering objects in our MFC model, we introduce the amplitude function as a stochastic process $a(\alpha, t)$ which flips on when it changes value from 0 to $a^+(\alpha) \geq 0$ and off when it changes values oppositely, cf. Figure 5.4. It is here assumed that the mapping $a^+ : \Omega \to \mathbb{R}_+$ is smooth; it might for example be constant or depend on the distance from scatterer to the receiver. The flip process is taken to be Poisson distributed with flip rate constant $C$:

$$P(a(\alpha) \text{ flips } k \text{ times on time step } \Delta t) = \frac{(C \Delta t)^k \exp(-C \Delta t)}{k!},$$

where flips are independent from the scatterers’ random initial phase shifts.

With the above defined amplitude function and the set of arrival angles $\{\alpha_j\}_{j=1}^M$ distributed according to a scatterer density $p(\alpha)$, we propose the following flip process extension of Clarke’s model

$$Z_{t,M} = \frac{1}{\sqrt{M}} \sum_{m=1}^M a(\alpha_m, t)e^{-i2\pi f_c v \cos(\alpha_m)t / c + \theta_m(t)}.$$  

(5.1.8)
CHAPTER 5. WIRELESS CHANNEL MODELING

Figure 5.3: **Left plot:** The autocorrelation function for Clarke’s model with azimuth density \( p(\alpha) = (2\pi)^{-1} \), \( v = 5\) m/s and \( f_c = 1.8775\) GHz. **Right plot:** The power spectral density of Clarke’s model, often called Jakes’ spectrum with the same model conditions as for the left plot.

Figure 5.4: Moving receiver in the center of a circular scattering environment with scatterers flipping on and off as time goes.

Realizations of (5.1.8) is then generated by generating scatterer angles \( \{\alpha_j\}_{j=1}^M \) and i.i.d. initial phase shifts \( \{\theta_m\}_{m=1}^M \), generating sample paths of the stochastic process amplitudes \( a(\alpha_m, t) \), and summing contributions according to (5.1.8). The left plot of Figure 5.5 illustrates the difference between an MFC signal realization envelope generated with \( C = 0 \) and one with positive flip rate. The positive flip
rate gives the signal envelope more small scale temporal noise and less smoothness than what is found in the non-flipping signal envelope. The right plot of Figure 5.5 is a measured signal envelope from Ericsson Labs which shows that the measured signal has a small scale noise contribution similar to the MFC signal realization with positive flip rate.

Figure 5.5: **Left plot:** Two computer generated signal realizations of the MFC model generated with same random seed initialization. Both realizations are generated with the modeling parameters $f_c = 1.8775\text{GHz}$, $v = 6.944\text{m/s}$, $a^+(\alpha) = 2/\sqrt{M}$, but the flip rate separates the realizations with the red dashed line corresponding to a realization having $C = 0$, and the whole line to a realization with $C = 5$. **Right plot:** Measured urban environment signal envelope from Ericsson Labs. The carrier frequency and receiver speed is identical to the corresponding values for the left plot.

### 5.2 From MFC Models to Gaussian Processes

In Paper I, we show that under some assumptions, the stochastic process $Z_{t,M}$ of the MFC model (5.1.8) converges to a Gaussian process as the number of included wave paths $M \to \infty$. Here we will give a short description of Gaussian processes, starting with the definition.

**Definition 5.2.1 (Gaussian Process)** A Gaussian process is a stochastic process $\{Z_t\}_{t \in [0,T], Z_t \in \mathbb{R}^n}$, for which any finite length sample vector $Z = (Z_{t_1}, Z_{t_2}, \ldots, Z_{t_n})$ with $0 \leq t_1 \leq t_2 \leq \ldots \leq t_n$ is multivariate normal distributed.

A numerical realization of a Gaussian process on a set of times $\{t_j\}_{j=1}^N$ can be created by first computing the process’ covariance matrix,

$$K_{i,j} = E[Z_{t_i}Z_{t_j}^*], \quad i, j \in \{1, 2, \ldots, N\}$$
and setting
\[ Z_{t_i} = \sum_{j=1}^{N} \sqrt{K_{i,j}} \chi_j, \quad (5.2.1) \]

where \( \sqrt{K} \) is the square root of \( K \) in the sense that it fulfills \( K = \sqrt{K} \sqrt{K}^H \) (for example the lower diagonal Cholesky factorization), and \( \chi_1, \chi_2, \ldots, \chi_N \) is a set of i.i.d. standard normal distributed r.v.

**Example 5.2.2 (The Wiener Process)** The Wiener process \( W_t \) is a Gaussian process which has the increment property \( W_{t_2} - W_{t_1} \sim N(0, |t_2 - t_1|) \) and thereby the covariance matrix
\[ K_{i,j} = E[W_{t_i} W_{t_j}] = E[W_{\min(t_i,t_j)}^2] = \min(t_i,t_j), \quad (5.2.2) \]
cf. Section 4.1. For the Wiener process, the structure of the Cholesky factorized \( \sqrt{K} \) is particularly simple:
\[
\sqrt{K} = \begin{pmatrix}
\sqrt{t_1} & 0 & \ldots & 0 \\
\sqrt{t_1} & \sqrt{t_2 - t_1} & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
\sqrt{t_1} & \sqrt{t_2 - t_1} & \ldots & \sqrt{t_N - t_{N-1}}
\end{pmatrix}.
\quad (5.2.3)
\]

This is fortunate because it makes the computational cost of generating a Wiener process realization \( \{W_{t_j}\}_{j=1}^{N} \) to \( O(N) \) which compares favorably to \( O(N^2) \) for general Gaussian process realizations. The Wiener process realization cost \( O(N) \) may be concluded from the scheme (5.2.1) which when \( \sqrt{K} \) is given by (5.2.3) becomes
\[ W_{t_{j+1}} = W_{t_j} + \sqrt{t_{j+1} - t_j} \chi_j, \quad \chi_j \sim \mathcal{N}(0,1). \]

Generating signal realizations from an MFC model is generally computationally costly, in particular in more realistic settings when a high number of wave paths \( M \) are included in your model. Comparatively, the analysis and computational experiments of Paper I indicate that generation of Gaussian process signal realizations are substantially less costly.
Chapter 6

Classical and Quantum Mechanics

For particles at human scale, motion is accurately described by classical mechanics, while for particles at atomic scale, motion is described by a quantum mechanics. In settings with many particles, quantum mechanical computations are very difficult, and, often, the best approximations one can make are by means of classical mechanics. Molecular dynamics is a classical mechanics approximation of the quantum scale motion of electrons and nuclei in molecular bindings, and in Paper IV of this thesis we study how well the motion of electrons and nuclei are approximated by molecular dynamics when the nucleus-to-electron mass ratio becomes large. In this chapter we give an outline of some classical and quantum mechanical concepts that will be useful when reading Paper IV.

6.1 Classical Mechanics

The motion of a single particle in a $d$-dimensional potential $V : \mathbb{R}^d \to \mathbb{R}$ and with mass $M$ is described by the force acting on the particle; $F = -\nabla V$. This yields the equations of motion

$$
\dot{q} = v \quad \text{and} \quad M\dot{v} = -\nabla V, \tag{6.1.1}
$$

where $q(t)$ and $v(t)$ represents the position and velocity of the particle, respectively. A preserved quantity during this motion is the energy $E = M|v|^2/2 + V(q)$ which is easily verified through differentiating:

$$
\dot{E} = M\dot{v} \cdot v + \nabla V \cdot \dot{q} = 0. \tag{6.1.2}
$$

Introducing the Hamiltonian $H(q, p) = |p|^2/2M + V(q)$ with the momentum $p = Mv$, it is often more convenient to represent the equations of motion on the Hamiltonian form

$$
\dot{q} = \nabla_p H \quad \text{and} \quad \dot{p} = -\nabla_q H. \tag{6.1.3}
$$

We also note that the Hamiltonian is preserved since by construction $H = E$. 

47
For a system of $N$ particles with the $i$th particle position denoted $q_i \in \mathbb{R}^d$ and with mass $M_i$, the force fields acting on each particle can often be divided into the external potential $V_i : \mathbb{R}^d \to \mathbb{R}$, and the potentials for the interaction between $i$th and $j$th particle $V_{ij}(q_i - q_j)$. Introducing the Hamiltonian

$$H(q, p) = \sum_{i=1}^{N} \frac{|p_i|^2}{2M_i} + V_i(q_i) + \sum_{i<j} V_{ij}(q_i - q_j),$$

where $q = (q_1, q_2, \ldots, q_N) \in \mathbb{R}^{dN}$ and, similarly, $p \in \mathbb{R}^{dN}$, the equations of motion for $N$ particles take the familiar form

$$\dot{q}_i = \nabla_p H \quad \text{and} \quad \dot{p}_i = -\nabla_q H, \quad \text{for} \quad i = 1, 2, \ldots, N. \quad (6.1.4)$$

The term particle is used in a relative sense in our above outline: we might be dealing with stars and planets or grains of sand, conditioned that the reduction to mass points makes sense. Let us present two examples of Hamiltonian dynamics.

**Example 6.1.1 (The harmonic oscillator)** The simple harmonic oscillator is 1-dimensional mass-spring system which when displaced from its equilibrium position experiences a restoring force $F = -Mkq$, with $k \in \mathbb{R}_+$ being a spring constant and $q \in \mathbb{R}$ the displacement from equilibrium. The Hamiltonian for this dynamics becomes

$$H(q, p) = \frac{p^2}{2M} + \frac{Mkq^2}{2}.$$ 

**Example 6.1.2 ($N$-body problem)** Newton explained the observed planetary motion by a negligible external field and interplanetary forces $F_{ij} = -GM_iM_j(q_i - q_j)/|q_i - q_j|^3$, where $G$ is a gravitational constant. Considering the setting with $N$ planets, Newton's interplanetary forces translate to the interaction potentials

$$V_{ij}(q_i - q_j) = -\frac{GM_iM_j}{|q_i - q_j|}, \quad i, j \in \{1, 2, \ldots, N\} \quad \text{and} \quad i \neq j,$$

and the negligible external potentials to $V_i = 0$, for $i = 1, 2, \ldots, N$. Consequently, the $N$-body equations of motion are implicitly given by (6.1.4) with the Hamiltonian

$$H(q, p) = \sum_{i=1}^{N} \frac{|p_i|^2}{2M_i} - \sum_{i<j} \frac{GM_iM_j}{|q_i - q_j|}.$$ 

**Evolution of States**

For a given $N$-particle system, the set of possible $(q, p)$ restricted to some subset of $\mathbb{R}^{dN} \times \mathbb{R}^{dN}$ is referred to as the phase space, and the vector $(q(t), p(t))$ at a given time $t$ is referred to as the particle system's state at time $t$. Considering experiments which involve measurements, we will extend the notion of states to
densities over the phase space \( \rho(q, p) \) with the restriction that \( \rho \) is non-negative. A pure state \((\tilde{q}, \tilde{p})\) may then be represented on density form by dirac functions \( \rho(q, p) = \delta(q - \tilde{q}, p - \tilde{p}) \). For a given state \( \rho \), the expected value of an observable, like the total energy \( H \), is given by

\[
E_{\rho}[H] = \frac{\int H(q, p) \rho(q, p) \, dq \, dp}{\int \rho(q, p) \, dq \, dp}.
\]

A convex combination of states also becomes a state. E.g., given two states \( \rho_1 \) and \( \rho_2 \), the convex combination \( \rho_3 = \alpha \rho_1 + (1 - \alpha) \rho_2 \) is also a state.

The motion of \((q, p)\) satisfies the Hamiltonian dynamics (6.1.4), and this implies that the density \( \rho(q, p) \) is convected by the flow \((\dot{q}, \dot{p})\). The convection makes it practical to consider the density time dependent in the following sense

\[
\rho(q(t), p(t); t) = \rho(q(0), p(0)). \tag{6.1.5}
\]

This is often referred to as the Liouville picture, cf. [15]. For any material region \( R(t) \) of phase space convected by the flow \((\dot{q}, \dot{p})\), the Liouville picture (6.1.5) straightforwardly implies that

\[
\frac{d}{dt} \int_{R(t)} \rho(q, p; t) \, dq \, dp = 0,
\]

and by analysis involving time differentiation of the Jacobian determinant of \( \partial(q, p) / \partial(q(0), p(0)) \), one may further derive the transport theorem

\[
\frac{d}{dt} \int_{R(t)} \rho(q, p; t) \, dq \, dp = \int_{R(t)} \partial_t \rho + \nabla_{q,p} \cdot ((\dot{q}, \dot{p}) \rho) \, dq \, dp, \tag{6.1.6}
\]

cf. [27]. For any fixed region \( D \), there is at any moment in time a coinciding material volume \( R(t) \), and combining this property with (6.1.5) and (6.1.6), yields the following density PDE

\[
\partial_t \rho + \nabla_{q,p} \cdot ((\dot{q}, \dot{p}) \rho) = 0.
\]

Furthermore, the Hamiltonian dynamics (6.1.4) implies that

\[
\nabla_{q,p} \cdot ((\dot{q}, \dot{p}) \rho) = \sum_{i=1}^{N} (\nabla_{q_i} \cdot \nabla_{p_i} - \nabla_{p_i} \cdot \nabla_{q_i}) H = 0, \tag{6.1.7}
\]

so that we end up with the following PDE for \( \rho(q, p; t) \)

\[
\partial_t \rho + \sum_{i=1}^{N} \nabla_{p_i} H \cdot \nabla_{q_i} \rho - \nabla_{q_i} H \cdot \nabla_{p_i} \rho = 0 \tag{6.1.8}
\]

with initial condition \( \rho(q, p; 0) = \rho(q, p) \). This is often referred to as Liouville’s theorem.
Preservation of Volume

For Hamiltonian dynamics, the phase space volume is preserved. This might be verified by considering the material region \( R(t) \) convected by the Hamiltonian flow \((\dot{q}, \dot{p})\) with the evolving phase space volume

\[
V(t) = \int_{R(t)} 1 \, dq \, dp.
\]

Equation (6.1.6), in this case with \( \rho = 1 \), and (6.1.7) implies that

\[
\dot{V}(t) = \int_{R(t)} \nabla_{q,p} \cdot (\dot{q}, \dot{p}) \, dq \, dp = 0.
\]

The preservation of phase space implies that if a fluid convected by Hamiltonian dynamics is expanding in \( q \)-space, then it is contracting in \( p \)-space, and vice versa.

Symplectic Numerical Methods

The word symplectic derives from the Ancient Greek \( συµπλεκτικός \) which is a composite of “braided” and “together”. In mathematics, it was introduced by Herman Weyl in 1939 to describe the group of unitary transformations of \( \mathbb{C}^{2n} \) that for vectors \( \xi = (\xi^q, \xi^p), \eta = (\eta^q, \eta^p) \) preserves the operation

\[
\sum_{i=1}^{n} (\xi_i^p)^* \eta_i^q - (\xi_i^q)^* \eta_i^p;
\]

where \( \xi^p, \xi^q, \eta^p, \eta^q \in \mathbb{C}^n \) are the \( p \) and \( q \) components of the vector. Introducing the matrix

\[
J = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix},
\]

with \( I \) the \( n \)-dimensional identity matrix, the preservation of the operation (6.1.9) is equivalent to the preservation of \( \xi^H J \eta \), and restricting ourselves to \( \mathbb{R}^{2n} \), we may with the aid of \( J \) define linear and differentiable symplectic maps as follows.

**Definition 6.1.3 (Symplectic linear map, cf. [19])** A linear map \( A : \mathbb{R}^{2n} \to \mathbb{R}^{2n} \) is called symplectic if

\[
A^T J A = J
\]

**Definition 6.1.4 (Symplectic differentiable map, cf. [19])** A differentiable map \( g : U \to \mathbb{R}^{2n} \), (where \( U \subset \mathbb{R}^{2n} \) is an open set) is called symplectic if the Jacobian matrix \( g'(q, p) \) is everywhere symplectic, i.e., if

\[
g'(q, p)^T J g'(q, p) = J.
\]
For any Hamiltonian and fixed \( t \), the flow mapping \( \phi_t(p(0), q(0)) = (p(t), q(t)) \) is symplectic, cf. [19]. This property opens an alternative proof of the flow \( \phi_t \): Symplecticity means that

\[
\phi_t'(q, p)^T J \phi_t'(q, p) = J.
\]

By observing that \( \det(J) = (-1)^{2n} = 1 \), \( \det(\phi_0'(q, p)) = 1 \) and assuming the Hamiltonian is sufficiently smooth to make \( \phi_t \) a continuous function with respect to \( t \), we conclude that \( \det(\phi_t'(q, p)) = 1 \) for all valid times. This implies conservation of volume.

Symplecticity is also sought property for numerical integrators of the Hamiltonian dynamics.

**Definition 6.1.5 (Symplectic one step-method, cf. [19])** A numerical one step method \( (p_{n+1}, q_{n+1}) = \phi_h((p_n, q_n)) \), is called symplectic if whenever applied to a smooth Hamiltonian system,

\[
\begin{pmatrix}
\frac{\partial(p_{n+1}, q_{n+1})}{\partial(p_n, q_n)} \\
\frac{\partial(p_{n+1}, q_{n+1})}{\partial(p_n, q_n)}
\end{pmatrix}^T J \begin{pmatrix}
\frac{\partial(p_{n+1}, q_{n+1})}{\partial(p_n, q_n)} \\
\frac{\partial(p_{n+1}, q_{n+1})}{\partial(p_n, q_n)}
\end{pmatrix} = J.
\]

For Hamiltonian systems, symplectic numerical integrators preserve the phase space volume perfectly and generally preserve energy approximately even when integrating over long time intervals. These properties are often highly important to preserve; in the case of planetary motion, for example, the preservation of the energy ensures that planets stay on orbit. Let us illustrate by an example comparing the performance simplest one step method, the forward Euler method

\[
(q_{n+1}, p_{n+1}) = (q_n + \Delta t p_n, p_n - \Delta t q_n),
\]

with the simplest symplectic one step method, the symplectic Euler method

\[
(p_{n+1}, q_{n+1}) = (q_n + \Delta t \partial_p H(p_{n+1}, q_n), p_n - \Delta t \partial_q H(p_{n+1}, q_n)).
\]

**Example 6.1.6 (Numerical solutions for the Harmonic oscillator)** We consider the Harmonic oscillator of Example 6.1.1 with \( M = 1, k = 1 \) and the Hamiltonian \( H = p^2/2 + q^2/2 \). For this problem, the forward Euler and symplectic Euler scheme take the following respective forms

\[
(q_{n+1}, p_{n+1}) = (q_n + \Delta t p_n, p_n - \Delta t q_n), \quad (\text{forward Euler}),
\]

\[
(q_{n+1}, p_{n+1}) = (q_n + \Delta t (p_n - \Delta t q_n), p_n - \Delta t q_n), \quad (\text{symplectic Euler}).
\]

With the initial condition \( (q(0), p(0)) = (1, 0) \), the preservation of the Hamiltonian implies that the dynamics should fulfill \( H(q, p) = 1/2 \) for any \( t \), and one may further derive that \( q(t) = \cos(t) \) is the solution of this problem. The comparison of the methods given in Figure 6.1 shows that while the volume is preserved and the energy is approximately preserved by the symplectic Euler method, even for large
time steps, the forward Euler method solutions have increasing material regions and energy.

![Numerical solutions of the Harmonic oscillator problem of Example 6.1.6](image)

Figure 6.1: Numerical solutions of the Harmonic oscillator problem of Example 6.1.6. The top four plots are numerical solutions with $\Delta t = 2\pi/32 \approx 0.2$ integrated on the interval $t \in [0, 2\pi)$ and consist of: phase space solution $(p(t), q(t))$ (upper left), $(t, q(t))$(upper right), the energy $(t, H(p(t), q(t))$ (lower left), anticlockwise flow simulation of phase space ball material regions (lower right). The bottom four plots are analogous numerical solutions with $\Delta t = 2\pi/64 \approx 0.1$. 

6.2 Quantum Mechanics

Quantum mechanics is a description of the mechanics of particles on atomic scale. Quantum mechanics differ from classical mechanics in the following sense:

- The state of a particle in classical mechanics is described by position and momentum (or, if extending notions, by a density) and the state evolves deterministically.

- The state of a particle in quantum mechanics is described by the amplitude of the complex-valued wave function \( \Phi(x,t) \) solving the time-dependent Schrödinger equation

\[
i\hbar \frac{\partial}{\partial t} \Phi = \left( -\frac{1}{2M} \Delta + V(x,t) \right) \Phi.
\]

Furthermore, since it is impossible to obtain perfect information on the state of a particle (an implication of Heisenberg’s uncertainty principle, cf. [18]), whether or not a quantum state evolves deterministically is a question never to be verified experimentally—probably.

Derivation of Schrödinger’s Equations

Around the beginning of the 1900s, various experiments, such as the the slit experiment(s), cf. [15], showed that elementary particles like photons and electrons in addition to having particle properties have wave-like diffraction properties. This wave-particle duality led to the conjecture that elementary particles both have particle attributes such as mass, charge, and spin, and wave attributes such as frequency and, consequently, that there are two equivalent expressions for the total energy. Classical mechanics states that the total energy for a single particle is given by

\[
E = \frac{|\mathbf{p}|^2}{2m} + V(x,t) \quad \text{with } \mathbf{p}, x \in \mathbb{R}^3.
\]

For waves, work by Planck, Einstein, and de Broglie led to the conjecture an elementary particle with momentum \( \mathbf{p} \) and energy \( E \) in some sense corresponds to a wave

\[
\Phi(x,t) = \phi(x,t)e^{i(k \cdot x - 2\pi ft)}
\]

with wave vector \( k \in \mathbb{R}^3 \) and frequency \( f \) for which the following relations hold

\[
\mathbf{p} = h\mathbf{k}, \quad |k| = 2\pi f/c, \quad \text{and} \quad E = |p|c = h|k|c = hf.
\]

Here \( c \) denotes the speed of light, \( h \approx 1.05 \times 10^{-27} \text{erg-sec} \) denotes Planck’s constant, and \( h := h/(2\pi) \). By the relations (6.2.2), the wave function (6.2.1) may be represented as follows

\[
\Phi(x,t) = \phi(x,t)e^{i(p \cdot x - Et)/\hbar}.
\]

Supposing the amplitude \( \phi \) is smooth and slowly varying compared to the scales \( \mathcal{O}(f/(hc)) \) and \( \mathcal{O}(E/h) \) the wave function (6.2.3) approximately fulfills the relation

\[
(i\hbar \partial_t + h^2 \frac{\Delta_x}{2m})\Phi = (E - \frac{|p|^2}{2m})\Phi.
\]
By coordinate transformations $t \rightarrow t/h, x \rightarrow x/h$ and replacing $E$ with the classical mechanics total energy $|p|^2/2m + V(x,t)$, equation (6.2.4) becomes the time-dependent Schrödinger equation

$$i\partial_t \Phi = \left( -\frac{1}{2m} \Delta + V(x,t) \right) \Phi.$$ 

Time-Independent Schrödinger Equation

Assuming the potential is time-independent, we may argue as above using the wave ansatz (6.2.3) to observe that $i\hbar \partial_t \Phi = E \Phi$ and to further derive the time-independent Schrödinger equation

$$\left( -\frac{1}{2m} \Delta + V(x) \right) \Phi = E \Phi.$$ 

Solutions of the time-independent Schrödinger equation describe particle motions with preserved total energy—steady state solutions. The time-independent Schrödinger equation requires no initial data $\Phi(\cdot,0)$, and in this sense it is a more fundamental PDE than the time-dependent Schrödinger equation. In what follows, we will restrict ourselves to the time-independent Schrödinger equation.

Many Particle Schrödinger Equation

Considering a molecule with $N$ nuclei and $n$ electrons, the time-independent Schrödinger equation takes the form

$$\left( -\sum_{i=1}^{N} \frac{1}{2M_i} \Delta_{X_i} - \sum_{i=1}^{n} \frac{1}{2m} \Delta_{x_i} + V(X,x) \right) \Phi(X,x) = E \Phi(X,x),$$

where $X = (X_1, X_2, \ldots, X_N) \in \mathbb{R}^{3N}$ and $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^{3n}$ denote the nuclear and electron positions, respectively, $M_i$ denotes the mass of the $i$-th nucleus, and $m \approx 9.11 \times 10^{-31} \text{kg}$ the electron mass. Supposing all nuclei have the same mass, i.e., $M_i = M_1 \forall i$, and introducing the coordinate transformations $(X,x) \rightarrow \sqrt{m}(X,x)$ and the short hand notation $\Delta_X := \sum_{i=1}^{N} \Delta_{X_i}$, we obtain the time-independent Schrödinger equation on the form studied in paper IV:

$$\left( -\frac{1}{2M} \Delta_X - \sum_{i=1}^{n} \frac{1}{2} \Delta_{x_i} + V(X,x) \right) \underbrace{\Phi(X,x)}_{=: \Phi(X,x)} = E \Phi(X,x),$$

with solutions sought in a Hilbert subspace of $L^2(\mathbb{R}^{3N} \times \mathbb{R}^{3n})$ with symmetry conditions based on the Pauli exclusion principle for bosons and fermions, cf. [23].
6.2. QUANTUM MECHANICS

In (6.2.5), \( M = \frac{M_1}{m} \) represents the nucleus-to-electron mass ratio which ranges from approximately 1836 for the \(^1\text{H}\) Hydrogen isotope\(^1\) (corresponding to the proton-to-electron mass ratio) to approximately \(244 \times 1836\) for the Uranium isotope \(^{244}\text{U}\), \(^{244}\text{U}\) being the heaviest known isotope to exist on earth in its natural form.

Quantum States

Each solution of the time-independent Schrödinger equation with energy \( E \in \mathbb{R} \) corresponds to a quantum state for the particle at the given energy level. That is, for any given solution with an energy \( E \), we assume the scaling

\[
\int \int |\Phi(X, x)|^2 \, dX \, dx = 1,
\]

and that \(|\Phi|^2\) is the probability distribution of the particle positions \((X, x)\) in the state \(\Phi\). For any observable, i.e., self-adjoint operators \(\mathcal{A}\) on \(L^{2}(\mathbb{R}^{3N} \times \mathbb{R}^{3n})\), the expected value is then given by

\[
E_{\Phi}[\mathcal{A}] = \int \int \Phi(X, x)^* \mathcal{A}(X) \Phi(X, x) \, dX \, dx.
\]

In paper IV, we consider the subset of observables consisting of self-adjoint operators on the nuclei coordinates \(L^2(dX)\). An example of an observable is the \(x\)-position of nucleus coordinate \(i\): \(\mathcal{A}(X) = X_{i,1}\).

For solutions of the time-independent Schrödinger equation a particular kind of linearity is observed: if \(\Phi_1\) and \(\Phi_2\) are solutions of (6.2.5) for a given energy \(E\) (so called degenerate solutions), then any linear combination of these solutions also solves (6.2.5) with the energy \(E\). That is,

\[
\left(-\sum_{i=1}^{N} \frac{1}{2M} \Delta x_i - V\right)(a_1 \Phi_1 + a_2 \Phi_2) = E(a_1 \Phi_1 + a_2 \Phi_2), \quad \forall a_1, a_2 \in \mathbb{C}
\]

The WKB Ansatz

In Paper IV, we compare the density generated by a state of the Schrödinger equation to a density generated by so called Born-Oppenheimer molecular dynamics. In what follows, we will give a short description of the mentioned densities, and some of the tools used to derive them, starting with the WKB ansatz.

The WKB ansatz assumes that solutions of the time-independent Schrödinger equation (6.2.5) are on the form

\[
\Phi(X, x) = \phi(X, x)e^{i\sqrt{\mathcal{M}} \theta(x)}, \quad (6.2.6)
\]

\(^1\)Isotope refers to the number of protons and neutrons in the nucleus of a chemical element.
where the amplitude $\phi : \mathbb{R}^{3N} \times \mathbb{R}^{3n} \rightarrow \mathbb{C}$ and the phase $\theta : \mathbb{R}^{3N} \rightarrow \mathbb{R}$ are smooth functions varying on a much slower scale than the mass ratio $M$. The WKB function is a solution of (6.2.5) provided that

$$0 = \left( -\frac{1}{2} \Delta \mathbf{x} + \mathcal{V}(\mathbf{X}, \mathbf{x}) - E \right) \phi(\mathbf{X}, \mathbf{x}) e^{i\sqrt{M} \theta(\mathbf{X})}$$

$$= \left( \frac{1}{2} |\nabla \theta|^2 + V_0 - E \right) \phi$$

$$= i \frac{1}{M^{1/2}} \left( \nabla \mathbf{x} \phi \cdot \nabla \mathbf{x} \theta + \frac{1}{2} \phi \Delta \mathbf{x} \theta \right) e^{i\sqrt{M} \theta(\mathbf{X})}.$$  

(6.2.7)

Here the term

$$V_0(\mathbf{X}) := \frac{\langle \phi(\mathbf{X}, \cdot), \mathcal{V}(\mathbf{X}, \cdot) \phi(\mathbf{X}, \cdot) \rangle}{\langle \phi(\mathbf{X}, \cdot), \phi(\mathbf{X}, \cdot) \rangle},$$

with $\langle \cdot, \cdot \rangle$ representing the electron coordinates inner product of complex-valued functions in $L^2(\mathbb{R}^{3n})$, is introduced so that setting $I = 0$ gives a well defined limit as $M \rightarrow \infty$ (see Paper IV, p. 9 for details). Setting $I = 0$ in (6.2.7) gives a (Eikonal) Hamilton-Jacobi PDE for the phase with characteristics given by the Hamiltonian system

$$\dot{\mathbf{X}} = \partial_P H_S, \quad \dot{\mathbf{P}} = -\partial_\mathbf{X} H_S, \quad \text{and} \quad H_S(\mathbf{X}, \mathbf{P}) = \frac{1}{2} |\mathbf{P}|^2 + V_0(\mathbf{X}) - E,$$

with $\mathbf{P}(t) = \nabla \mathbf{x} \theta(\mathbf{X}(t))$. Thereafter, setting $II = 0$ in (6.2.7), we derive in Theorem 3.1 of Paper IV that under some conditions, the phase fulfills the equation

$$\phi(\mathbf{X}(t), \mathbf{x}) = \frac{\psi(\mathbf{X}(t), \mathbf{x})}{G(\mathbf{X}(t))},$$

where $\psi(\mathbf{X}, \mathbf{x})$ is the solution of the time-dependent Schrödinger equation

$$iM^{-1/2} \frac{d}{dt} \psi(\mathbf{X}(t), \mathbf{x}) = \left( \mathcal{V}(\mathbf{X}(t), \mathbf{x}) - V_0(\mathbf{X}(t)) \right) \psi(\mathbf{X}(t), \mathbf{x}) - \frac{G(\mathbf{X}(t))}{2M} \Delta \mathbf{x} \frac{\psi(\mathbf{X}(t), \mathbf{x})}{G(\mathbf{X}(t))},$$

and $G : \mathbb{R}^{3N} \rightarrow \mathbb{R}$ is implicitly defined by the integrating factor

$$\frac{d}{dt} \log G(\mathbf{X}(t)) = \frac{1}{2} \Delta \theta(\mathbf{X}(t)).$$

Supposing the quantum state generated from the above equations for the phase and amplitude is well defined, the WKB ansatz (6.2.6) takes the form

$$\Phi(\mathbf{X}, \mathbf{x}) = \phi(\mathbf{X}, \mathbf{x}) e^{i\sqrt{M} \theta(\mathbf{X})} = \frac{\psi(\mathbf{X}, \mathbf{x})}{G(\mathbf{X})} e^{i\sqrt{M} \theta(\mathbf{X})}.$$
and the nuclear coordinate density becomes

\[ \rho_S(X) = \frac{G^{-2}(X) \langle \psi(X, \cdot), \psi(X, \cdot) \rangle}{\| G^{-1} \psi \|_{L^2(\mathbb{R}^{3N} \times \mathbb{R}^{3n})}}. \]

We next present an approximation of the density \( \rho_S \) generated by Born-Oppenheimer molecular dynamics.

The Born-Oppenheimer Approximation

The Born-Oppenheimer approximation of solutions of the time-independent Schrödinger equation (6.2.5) consists of the following two steps:

1. Clamp the nuclear coordinate and neglect the nuclear kinetic energy in the equation (6.2.5) and solve the remaining electron coordinate equation

\[ V(X, \cdot) \psi_0(X, \cdot) = \lambda_0(X) \psi_0(X, \cdot). \]  \hspace{1cm} (6.2.8)

Here \( \lambda_0(X) \) represents fixed eigenvalue function\(^2\) of the operator \( V(X, \cdot) \) which we assume is spectrally separated from other eigenvalues. This equation is generally solved approximately.

2. Reintroduce the kinetic energy into the equation and approximately solve the nuclear coordinate equation

\[ \left( \frac{1}{2M} \Delta X + \lambda_0(X) \right) \Phi(X) = E \Phi(X). \]  \hspace{1cm} (6.2.9)

The separation of nuclear and electron coordinates in the Born-Oppenheimer approximation is motivated from the assumption that the relatively speaking heavy nuclei move at a much slower speed than the light electrons (a valid assumption if nuclear and electron momenta are of comparable magnitude).

To motivate the Born-Oppenheimer molecular dynamics we first state the ansatz

\[ \Phi_{BO}(X, x) = \phi_{BO}(X, x) e^{i \sqrt{M} \theta_{BO}(X)}, \]

where, as before, we assume that the amplitude \( \phi_{BO} : \mathbb{R}^{3N} \times \mathbb{R}^{3n} \rightarrow \mathbb{C} \) and phase \( \theta_{BO} : \mathbb{R}^{3N} \rightarrow \mathbb{R} \) are smooth functions which vary on a much slower scale than \( M \). To conform with step 1. of the Born-Oppenheimer approximation, we assume that the electrons are in an eigenstate \( \psi_0(X, \cdot) \) of (6.2.8) with eigenfunction \( \lambda_0(X) \) so that the amplitude takes the form \( \phi_{BO}(X, x) = \sqrt{\rho_{BO}(X)} \psi_0(X, x) \) with unknown

\(^2\)Often referred to as a potential energy surface.
density function \( \rho_{BO} : \mathbb{R}^{3N} \rightarrow \mathbb{R} \). Inserting \( \Phi_{BO} \) into equation (6.2.7), yields

\[
0 = \left( \frac{1}{2M} \Delta X + \lambda_0(X) - E \right) \Phi_{BO}
\]

\[
= \left( \frac{1}{2} |\nabla \theta_{BO}|^2 + \lambda_0 - E \right) \phi_{BO}
\]

\[
- \frac{1}{2M} \Delta X \phi_{BO} - \frac{i}{M^{1/2}} (\nabla X \phi_{BO} \cdot \nabla \theta_{BO} + \frac{1}{2} \phi_{BO} \Delta \theta_{BO}) \right) e^{i\sqrt{M}\theta_{BO}(X)}.
\]

(6.2.10)

We approximately solve this equation by truncating the \( \mathcal{O}(M^{-1/2}) \) terms and considering the remaining the Eikonal equation of term I:

\[
\frac{1}{2} |\nabla \theta_{BO}(X)|^2 + \lambda_0(X) - E = 0,
\]

whose characteristics are given by the Hamiltonian system

\[
\dot{X}_i = \partial_{P_i} H_{BO}(X, P), \quad \dot{P} = -\partial_{X_i} H_{BO}(X, P),
\]

where \( H_{BO}(X, P) := \sum_{i=1}^{N} \frac{|P_i|^2}{2M} + \lambda_0(X) \), and \( P(t) = \nabla X \theta_{BO}(X(t)) \).

(6.2.11)

To determine the density \( \rho_{BO} \) for the nuclear coordinates described by the dynamics (6.2.11), we first observe that conservation of mass implies that

\[
\nabla \cdot (\rho_{BO}(X) \nabla \theta_{BO}(X)) = 0.
\]

By the conservation of mass and the dynamics (6.2.11) we derive the ordinary differential equation

\[
\frac{d}{dt} \rho_{BO}(X(t)) = \nabla \rho_{BO}(X(t)) \cdot \nabla \theta_{BO}(X(t)) = -\rho_{BO}(X(t)) \Delta \theta_{BO}(X(t)).
\]

Introducing the integrating factor

\[
\frac{d}{dt} \log G_{BO}(X(t)) = \frac{1}{2} \Delta \theta_{BO}(X(t)),
\]

we obtain the following relation for the Born-Oppenheimer molecular dynamics density

\[
\rho_{BO}(X) = \frac{C}{G_{BO}^2(X)}.
\]
The main study of Paper IV is the comparison of the densities
\[ \rho_S(X) = \frac{G^{-2}(X)\langle \psi(X, \cdot), \psi(X, \cdot) \rangle}{\|G^{-1}\psi\|_{L^2(\mathbb{R}^3N \times \mathbb{R}^3n)}} \quad \text{and} \quad \rho_{BO}(X) = \frac{C}{G_{BO}^2(X)}, \]
and we end this chapter by noting that Theorem 7.1 of Paper IV we prove that under some assumptions
\[ \int_{\mathbb{R}^3N} g(X)\rho_{BO}(X) dX = \int_{\mathbb{R}^3N} g(X)\rho_S(X) dX + O(M^{-1+\delta}), \quad (6.2.12) \]
for any \( \delta > 0 \) and \( g \in C^3(\mathbb{R}^{3N}) \). The proof of (6.2.12) relies strongly on the stability of symplectic numerical methods for Hamiltonian-Jacobi equations derived by Szepessy et al. in [29].
Bibliography


